2015

Development and validation of virtual interactive tasks for an aviation English assessment

Moonyoung Park
Iowa State University

Development and validation of virtual interactive tasks for an aviation English assessment

by

Moonyoung Park

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Applied Linguistics and Technology

Program of Study Committee:
Carol A. Chapelle, Major Professor
Dan Douglas
Gary Ockey
Geoffrey Sauer
Ana Correia
Randall Sadler

Iowa State University
Ames, Iowa
2015

TABLE OF CONTENTS

Page

Page

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude to my committee chair and major professor, Dr. Carol Chapelle, for her support and guidance throughout the research. Her infinite professional insight and breadth of knowledge have encouraged me towards the completion of this dissertation. I would also like to extend my appreciation to my committee members, Dr. Dan Douglas, Dr. Gary Ockey, Dr. Geoffrey Sauer, Dr. Ana Correia, and Dr. Randall Sadler, for their time, support, motivation, and help whenever necessary in every step of my academic journey.

My sincere appreciation is extended to the military air traffic controllers in Korea for expressing their enthusiasm for and willingness to participate in this research. Very special thanks go to CSM Jewon Ryu in the 55th Air Traffic Service Battalion, who helped me in various ways in both the pilot study and dissertation study. I also extend my thanks to Dr. Byeong-Young Cho at University of Pittsburgh for his friendship and insights into the world of strategic competences and Dr. Alistair Van Moere at Pearson for providing me with Versant Aviation English Tests for this dissertation research.

I would also like to acknowledge my PhD colleagues and faculty and staff in the English Department at Iowa State University for their continued support, care, and collaboration. I would further like to thank my colleagues in the Intensive English and Orientation Program (IEOP) at ISU and the Center for Language Research at University of Aizu in Japan for their unwavering encouragement and generous support.

My heartfelt thanks go to my sister's family − Seoyoung, Joonyoung, and Dongjoo − for their love, support, and laughter when I needed it most. I am deeply indebted to my parents who taught me the value of hard work and perseverance, not to mention their love and prayers.

This dissertation is dedicated to my two professors at Keimyung University in Korea, Richard Hark and Yongsang Cho, who had been my greatest inspirations and role models since 1996, but sadly passed away soon after I began the doctoral program. I believe they would be most pleased to see the completion of my PhD journey.

ABSTRACT

In response to growing concerns over aviation safety stemming from the limited command of aviation English by non-native English speaking practitioners, this study aimed to demonstrate the development process of aviation English test tasks in a virtual environment and investigate the validity for a task-based aviation English performance assessment in the context of Korean Army Aviation. In the current dissertation study, the development and validation of the VITAEA test were based on four inferences – domain description, evaluation, generalization, and explanation – and underlying assumptions in an interpretive argument that developed with reference to argument-based validity, evidence-centered design, target language use situation analysis for test development in language for specific purposes, and task-based language assessment. Adopting a mixed method with a triangulation design, qualitative and quantitative evidence was collected to provide valid support for the inferences and to strengthen the validity argument.

Based on a task-based needs analysis with 81 military air traffic controllers on the required aviation English knowledge, skills, processes, target tasks, and task procedures in the TLU situations, virtual interactive aviation English tasks were developed in Second Life. A total of 20 controllers completed the prototype virtual interactive tasks for aviation English assessment, and their output was then rated by two rater groups, one engaging in task-centered rating and one accomplishing language-centered rating. Data included 20 task-based performance assessment sample audio files; 19 follow-up test taker interviews and online survey questionnaires; three language-centered raters' post-rating questionnaire responses; two task-centered raters' post-rating interview transcripts; military aviation English training manuals and references; and coded transcripts of 12 test takers' stimulated recall and their actual task performance.

The validity evidence collected in the various phases of test development and validation serves as backing for the four inferences in the interpretive argument as well as provides invaluable resources for the revision of the prototype virtual interactive tasks for aviation English assessment. Furthermore, empirical processes for prototype test development and partial validation based on the theoretical guidance presented in this dissertation study can be seen as among the first to be constructed utilizing the three theoretical frameworks – argument-based approach, evidence-centered design, and task-based language assessment. In addition, this dissertation study can shed light on the steps required in application of an argument-based approach for task-based second language assessment. Lastly, this study provides additional grounds for the potential use of an immersive interface and simulated target language use situations in virtual environments to provide test takers with more authentic opportunities to perform the target tasks.

CHAPTER 1

INTRODUCTION

Aviation English, a key component of air traffic control communication, can be defined as a comprehensive but specialized subset of English for Specific Purposes (ESP) related broadly to aviation and consisting of both plain language and standardized phraseology for radiotelephony communications. ICAO recently recognizes the importance of plain language, especially in non-standard situations (e.g., referring to a violet passenger and describing failed landing gear), where standard phraseology is employed (Gerighty, 2008). Aviation English includes the use of English relating to any aspect of aviation by maintenance technicians, flight attendants, dispatchers, or managers and officials within the aviation industry (ICAO, 2004).

Notable characteristics of aviation English between air traffic controllers and pilots are that it is used in highly predictable circumstances, and normal communications follow a prescribed sequence (Mell, 1992). Unfortunately, however, even in such predictable and restricted circumstances, miscommunications can and do occur, especially between non-native English speaking pilots and air traffic controllers, resultant from a number of factors. These factors may include pilots not realizing a specific communication is intended for them, interference on the radio frequency, overlapping calls, misunderstood flight parameters, incorrect read backs, and inadequate clarification of flight parameters, to name a few (Cushing, 1994).

Researchers who are exploring language use in air traffic control (ATC) performance have tended to focus on the limitations of non-native English speakers and associated threats for aviation safety due to their limited command of aviation English (Atsushi, 2003, 2004). In line with the empirical research, recent International Civil Aviation Organization (ICAO) policy requires non-native English speaking pilots, navigators, air traffic controllers, and station

operators to demonstrate their aviation English ability to speak and understand the language used for radiotelephony communications. National Aeronautics and Space Administration (NASA) (1996) also acknowledges the problem of non-native English speakers' ATC performance, reporting that "25% (out of 28,000) of the reports cited a language problem as a primary cause of the foreign airspace operational incidents reported to the Aviation Safety Reporting System (ASRS)."

Following researchers' recommendations, the ICAO has developed a set of language proficiency requirements (LPRs), which consist of six levels of skills in six areas of language use: pronunciation, structure, vocabulary, fluency, comprehension, and interaction. According to the policy, the ICAO demands non-native English speaking pilots, navigators, air traffic controllers, and station operators to demonstrate their aviation English ability to speak and understand the language used for radiotelephony communications. However, recent studies (Alderson, 2009; Atsushi, 2003, 2004) conclude that many of the aviation English assessment procedures do not appear to meet international professional standards for language tests. Thus, the implementation of the language assessment policy is inadequate, and much more careful and close monitoring of the quality of the tests and assessment procedures is needed.

Contrary to civil aviation, military aviation contexts present particularly difficult study environments for researchers investigating the teaching and assessment of aviation English in some countries, a consequence of governmental restrictions and security issues; Republic of Korea is one such country where constraints of the military aviation context have prohibited research into practices in aviation English assessment. According to Ryu (2012, 2013), the conventional aviation English test in Republic of Korea (ROK) military aviation, the context for this dissertation study, is not free from such issues of inadequate aviation English assessment.

Acknowledging the importance of testing Korean air traffic controllers' aviation English ability for over a decade, the two concerned authorities, the Army Aviation School and an Air Traffic Services (ATS) Battalion in the Army Air Operations Command in Republic of Korea, have used conventional aviation English tests, a multiple choice, grammar, reading, and vocabulary-focused aviation English test that measures novice and experienced air traffic controllers' aviation English proficiency. Though the conventional Korean Army aviation English test has been inadequate in meeting the needs of the test users and international professional standards for language tests, the primary use of the aviation English test is to appropriately distribute throughout the Korean army bases the novice and experienced air traffic controllers with respect to their demonstrated level of aviation English communication skills. Based on the test results, those who score higher are considered capable of successfully accomplishing particular real-world tasks and being deployed in a more strategic flight coordination center (FCC), a ground control center (GC), or an airbase tower (TWR). Conversely, those who score lower are sent to the ground control center or airbase tower, where soldiers' positions and duties are of lesser strategic importance and there is a lighter workload. The secondary use of the assessment, and one in line with the intended purposes of the test, is to provide diagnostic feedback to the air traffic controllers in order to improve their aviation English communication skills.

The issues identified in this brief review of the conventional Korean Army aviation English test, as well as other civil aviation English tests [e.g., Test of English Language Proficiency for Aeronautical Communication (ELPAC) by EUROCONTROL (Alderson, 2009); Test of English Language Level for Controllers and Pilots (TELLCAP) by TELLCAP (Alderson, 2006); and RMIT English Language Test for Aviation (RELTA) by RMIT University (Zokić,

Boras, & Lazić, 2012)], are significant for a number of reasons. First and foremost, the conventional aviation English test in Republic of Korea is precisely the case of construct under-representation, which could present major threats to the validity of any test (Messick, 1989). The current Korean army aviation English test tasks measure test takers' aviation English grammar, reading, and vocabulary with paper-based multiple choice test items. The given facets of the construct of aviation English proficiency appear to be poorly defined and fail to include test tasks that enable raters to observe and evaluate test takers' aviation English performance of actual listening and speaking, highly significant dimensions of the construct of authentic aviation English communication.

Second, despite the promotion of the standard quality of the tests with regards to ICAO's six LRPs, little or no confidence could be placed in the reliability and validity of the several currently available aviation language tests (Alderson, 2006). This dissertation study researcher's hands-on experience with sample aviation language test items indicates that the majority of tests have a static, one-way task construction with little or no interaction during the item or task completion making the test far from representative of authentic, real-life ATC situations. Although most civil aviation English testing companies promote their online test service features, they do not make full use of the online testing environment, but simply present motionless pictures and plain text prompts to play audio materials and record test takers' voices instead of integrating interactive features and assessing ATC task performance. This being the case, recent efforts of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) to develop prototype interactive training applications and assessments in maturing learning technologies, including virtual environments and collaborative game-based technologies, seem to be highly thought-provoking and refreshingly positive (Brusso, Wisher,

Paddock, & Hatfield, 2014). Such endeavors are very much in line with the latest trend of assessment in K-12 classrooms in the U.S. to enhance the feasibility of virtual performance assessments and evaluate scientific inquiry by providing authentic environments (Clarke-Midura & Dede, 2010).

Third, even though the set of LPR's assessment criteria developed by ICAO were well-defined and appropriate for assessing complex and multidimensional aviation English proficiency, such a language-centered assessment rubric would still be unable to directly determine successful accomplishment of the particular target tasks. Because evaluating proficiency is considered the primary use of the Korean army aviation English test, the task-based performance assessment needs to be both construct-centered and task-centered (Messick, 1994), viewing a test task as a vehicle for determining task accomplishment and the elicitation of language performance, respectively. From this use perspective, the criteria for task-based aviation English performance assessment should not only inform inferences about test takers' aviation English proficiency (ability), but also provide test users with the results of successful completion of authentic tasks.

One of the primary impetuses for choosing army aviation English assessment as the context for this dissertation study can be attributed to the researcher's own military ATC training and work experience in the 55th ATS Battalion from 1997 to 1999 where, during his military service, the researcher noticed a discrepancy between the assessment and actual practice of ATC communication. Another impetus derives from the concerns raised during communication with Command Sergeant Major (CSM) Ryu, a personal acquaintance of the researcher who has been in charge of ATC training and assessment in the battalion. Both CSM Ryu and the researcher witnessed and discussed many problems regarding the aviation English assessment and training

during their work together in the 55th ATS Battalion. This dissertation study aims to uncover issues in the development and implementation of a prototype virtual interactive tasks for aviation English assessment, thereby contributing to the field of ATC English instruction and assessment in both the Korean military context and possibly in air traffic control communication research as a whole.

Research Goals

Given the diverse expected uses of the Korean Army aviation English test, including decision-making for controllers' placement assignment and diagnostic feedback for supplementary ATC training, its serious issues of construct under-representation and lack of authenticity, and growing potential of task-based performance assessment in virtual environments, the research goals for the dissertation study are to develop a prototype of virtual interactive aviation English test (VIAET) and to demonstrate test validation in which claims about test score interpretations and use are supported by assumptions and backing under the framework of interpretive argument (Chapelle, 2008; Chapelle, Enright, & Jamieson, 2008; Kane, 2006). The aspects the researcher pays particular attention to are military air traffic controllers' perceptions of and test-taking processes in the virtual interactive testing environment.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews previous research studies in four areas that serve as the foundation of the current study: (a) Language (English) for specific purposes (LSP/ESP) assessment, (b) aviation English, (c) task-based performance assessment, and (d) virtual worlds (Second Life). The section begins with ESP assessment which ties all of the following concepts together followed by the definition of aviation English, a subset of ESP, including its characteristics and the context of aviation English communication for this study. Next, task-based performance assessment is introduced with regards to the synergistic effect of implementing it in a virtual world. Then, the advantage and potential of a virtual environment for language assessment is introduced as the platform for assessment development. This chapter ends with a list of research issues for the current dissertation study.

Language for Specific Purpose

The main research domain of this dissertation study on aviation English assessment in a virtual environment is testing in Language for Specific Purposes (LSP). Reflecting the goal of the general purpose of language teaching as the development of learners' communicative capacity to achieve diverse communicative goals, the goal of general language tests is mainly to assess examinees' communicative capacity with limited or no substantial reference to the language use situations (Widdowson, 1983). However, LSP testing aims to measure the ability (or abilities) to manipulate language functions appropriately in a specific area of language use, namely English for Academic Purposes (EAP) and English for Employment Purposes (EEP) (e.g., aviation English, business English, medical English), in a variety of ways (Davies, 2001;

Lomperis, 1996). Testing LSP can be defined as a branch of language testing of which the test contents and methods are based on a target language use (TLU) analysis and the purpose of LSP testing. Although the definitions of general language tests and specific purposes tests may sound overlapping to some extent, LSP tests can be distinguished from general language tests in the following two aspects: (1) authenticity of task and (2) the interaction between language knowledge and specific purpose content knowledge (Douglas, 2000). Among diverse TLU settings, the target domain of this dissertation study is aviation English. Douglas's insights on the authentic task development and the interaction between language knowledge and content / background knowledge inspired the development of interactive aviation English tasks for this dissertation study.

Pedagogy and assessment in ESP as a coherent field of study has a relatively short history (Swales, 1985). The establishment of the University of Cambridge Local Examinations Syndicate's (UCLES) Certificate of Proficiency in 1913, which aimed to measure prospective English teachers' language proficiency, may be one of the earliest cases of LSP testing. Another instance of early LSP testing was the College Entrance Examination Board's English Competence examination in the U.S. in 1930, a test which measured international applicants' language ability in the context of U.S. college and universities (Douglas, 2000). However, despite the fact these two tests clearly aimed to assess vocational and academic English, the tests were lacking in the analysis of target language use (TLU) situation, which means a set of specific language use tasks that the test taker is likely to encounter outside of the test itself (Bachman & Palmer, 1996). Such a limited analysis of TLU led to restricted authenticity.

About five decades ago, production of ESP course materials for basic scientific English (Ewer, Latorre, & Derneği, 1969) and technical English (Herbert, 1965) based on register-based

studies (Barber, 1962) gained increased attention. The primary focus of ESP pedagogy at that time was placed on a sentence-bound view of language based on the frequency of certain syntactic forms and lexical items to prioritize the content of related course syllabi (Skehan, 1984). More recently, the shift in emphasis towards the discourse level and beyond the sentence-bound view was employed to perceive language as a functional system and not as isolated elements (Robinson, 1980). One major outcome of this trend was a book, "Communicative syllabus design" (Munby, 1978) which adopted a needs analysis as the basis for syllabus specification and later influenced the development of some new ESP tests (Carroll, 1980).

Munby employed the notion of a communicative syllabus design which consists of two stages: (1) a structured information gathering procedure for a course writer to develop a detailed special purpose syllabus as input for the next stage (communicative needs processor) and (2) converting the needs into syllabus specification. Such an approach potentially provided considerable support to ESP test developers by providing a well-defined set of categories on test content (Skehan, 1984). Contrary to this trend, however, groups of researchers attempted to demonstrate that there exists one underlying competence which can explain language performance (Oller, 1976; Oller & Perkins, 1980). Their concern was to find a test or test battery which can be the best single measure of unitary competence.

One exemplary LSP test, which seems to reflect the voice from the two trends, is the Temporary Registration Assessment Board (TRAB) examination developed in 1975 by the British General Medical Council to evaluate the professional and language abilities of physicians who were trained outside the United Kingdom. The examination is structured to measure professional competence as well as communication ability in English based on an analytical approach to and in-depth analysis of both spoken and written languages used by physicians,

nurses, and patients in the target contexts. The test consists of (1) a language component assessment through a taped, listening test and a written essay, and (2) both a professional knowledge and language ability assessment through an oral interview. TRAB developers aimed to provide appropriate and contextually authentic features in the test materials so that they could engage test takers' language ability as well as background knowledge in the test tasks (Douglas, 2000). Rea-Dickins (1987) also highlighted the collaborative efforts between language testing experts and medical experts to construct the TRAB? as the pre-requisite for the special purposes test design. The case of TRAM, which highlighted the importance of on-going collaboration between test developers and target domain experts, actually provides insights for this dissertation study so that the prototype aviation tasks could provide examinees with a more authentic simulation environment.

Despite ongoing efforts to develop and implement more appropriate LSP tests, skepticism of LSP testing was expressed by a group of later researchers. Davies (2001) points out a boundary problem of target languages in LSP. The first issue is a theoretical one – does English for Specific Purposes exist as a theoretical construct? – while the second issue is a practical one – how specific is *specific*? He took an example of 'Medical English' as an overlap with 'chemical English', on the one hand; on the other, 'medical English' itself consists of English of surgery, general medicine, pediatrics, and so on. At the lowest level, Davies suggests that an LSP equals the target language when used for a particular communicative purpose. To some extent, it is true that some languages from different LSP tests are overlapping to each other like the case of medical English and chemical English. Aviation English also shares much jargon in common with nautical English, as they both are based on radio phraseology in the language level. However, the identified issue of conventional LSP testing addressed above is that the assessment

mainly lies in target language level with little consideration what test takers can actually perform with the target language in an authentic test environment (Douglas, 2000). Accordingly, it is critical to consider interpreting test takers' task performance as evidence of their language ability in a TLU situation.

Another issue in the LSP especially in test design and validation is authenticity. This unitary notion of authenticity encompasses two aspects of authenticity: situational and interactional. Situational authenticity focuses on a relationship between an assessment task and the corresponding real-life situation, while interactional authenticity is involved in a test taker's language ability in assessment task completion (Bachman, 1990). As Douglas (2000) highlights, the two aspects of authenticity are the primary factors distinguishing language tests for specific purposes from those for relatively general purposes. If a test task is closely connected to tasks in a specific context, performance on the test task can be inferred as a test taker's ability in the real-world context of that specific purpose. Thus, assessing language for specific purposes is involved with interwoven characteristics of test takers' language ability, assessment tasks, and TLU situations.

As the concept of authenticity is rather subjective and is challenging to quantify, for the current dissertation research, the researcher adopts qualitative data from the TLU situations of a military air traffic control tower in the Korean Army Aviation. The first authenticity aspect of situation is concerned with authentic characteristics derived from an analysis of tasks in the TLU situation. This is verified by the test takers and test users who currently serve as air traffic controllers in the Korean Army Aviation. The second authenticity aspect of interaction is closely related to interaction of the test taker's aviation English ability with the test task. To address this

second concept of authenticity, the researcher adopts a post-test interview and a verbal protocol method (stimulated recall) with test takers right after the task-based aviation English test.

Lastly, criticism of LSP testing is concerned with content representativeness issue of selecting representative test tasks from the behavioral domain of interest (i.e., TLU situations). Content representativeness refers to the extent to which a test samples the content domain of interest (Bachman, 2002). The issue of investigating and demonstrating content representativeness is two-folds – identification of the TLU domain that the test-taker is likely to encounter outside the test itself and selecting test tasks from the domain (Bachman & Palmer, 1996). Unfortunately, however, it is challenging to provide evidence for content representativeness if the TLU domain is not clear, or is ill-defined. This could be true especially in the case of English for Academic Purposes (EAP), as its TLU is quite broad and challenging to narrow down to an ideal number of test tasks for EAP tests. To address this issue of selecting tasks from the TLU domain for test task development, needs analysis is recommended by numerous studies (Long, 1985; Bachman and Palmer, 1996; Norris et al., 1998).

In fact, aviation English, the target language of this dissertation study, includes a much smaller number of vocabulary items for its domain, and its TLU situations could also be categorized and sorted out in a relatively clear way. Furthermore, for the current dissertation study, the TLU situation of the LSP test development is intentionally limited to aviation English used in the air traffic control (ATC) tower context of army aviation expecting more fine-grained test task identification and selection. Further introduction about aviation English will be provided in the following section.

Aviation English

A subdomain of LSP and the target language domain of this dissertation study is Korean army aviation English for air traffic controllers. This section first reviews the definition of aviation English and its characteristics, specifies army aviation English in Korean contexts, and provides a construct definition of aviation English. The section concludes with a review of the aviation English assessment studies that provided the background for current dissertation study.

Definition of Aviation English

Aviation English is a subset of English for Specific Purposes (ESP) related broadly to aviation, and consists of both the plain language and aviation phraseologies for radiotelephony communications. Douglas (2004) highlights that it is necessary to identify the nature of aviation English, both the standardized phraseology and plain English, the relationship between them, and the situations when each is adopted. Research (Howard, 2008; Kim & Elder, 2009) has shown that plain English tends to be adopted when speakers are involved in abnormal or emergency circumstances, even when aviation phraseology could suffice. Such underutilization of aviation phraseology results in more complex use of structure and vocabulary and could prompt a communication problem between speakers.

Aviation English includes the use of English relating to any aspect of aviation by maintenance technicians, flight attendants, dispatchers, or managers and officials within the aviation industry. For example, aviation English could contain the language needed by pilots or air traffic controllers for briefing, making announcements, and communicating on the flight deck (ICAO, 2004). In this study, the focus of aviation English is restricted to communications between air traffic controllers and pilots.

Standards and Recommended Practices (SARPs) for Aeronautical Telecommunications were first adopted by the Council on May 30th 1949, pursuant to the provisions of Article 37 of the Convention on International Civil Aviation (Chicago 1944) and designated as Annex 10 to the Convention. The basic aviation phraseology principles have much in common with the underlying conventions of a controlled natural language based on English and designed to facilitate communication between ships (Seaspeak) (Kim, 2012). Aviation and nautical phraseologies have evolved over time with periodic initiatives by concerned organizations to codify and standardize their use to provide maximum clarity and brevity, and eliminate ambiguity in communications.

The language system of the phraseologies is the English-based radiotelephony system, as English is accepted as the lingua franca of aviation. This raises an important issuel researchers (e.g., Douglas, 2014; Kim & Elder, 2009; McNamara, 2012) have argued recently that aviation English is better thought of as a lingua franca than as ESP. Research on English for lingua franca also has criticized the conventional belief that native speakers own English and non-native English speakers should speak English following the standards of the native speakers (Jenkins, 2000). As highlighted by Kim and Elder (2009), aviation communication problems should be attributed to both native and non-native speakers. This critical trend is reflected in the second edition of ICAO DOC 9835, which claims that “the burden of improved communication should not be seen as fallen solely on non-native speakers” (Section 5.3).

Characteristics of Aviation English

Aircrafts in controlled airspace are all required to follow certain procedures and maneuver using corresponding phrases which are documented and approved by aviation

authorities. Such phrases constitute the language of routine air-ground communication and are known as standard radiotelephony phraseology. Aviation English for international air traffic control adopts a prescribed code of predetermined phrases with non-idiomatic forms and usage to avoid ambiguity between pilots and air traffic controllers. Due to the restricted, repetitive, and situationally dependent nature of radiotelephony phraseology, all pilots and air traffic controllers are trained to memorize such standardized phraseology in English (Emery, 2014).

Aviation English has its own unique International Radiotelephony Spelling Alphabet. The phonetic alphabet employed in aviation English is used to spell out letters instead of saying the letter itself in order to avoid listener confusion. For example, during the radio transmission, some letters such as “B” and “D” can be easily misunderstood, as both consonants end with the same /i/ vowel sound. Using the Phonetic Alphabet (see Table 1), they can be clearly distinguished with the code words “Bravo” and “Delta.” The following alphabet substitutes an entire word to represent one letter. The initial letter of each word is the letter of the alphabet it stands for.

Table 1.

ICAO Phonetic Alphabet

Letter	**Code Word:**	**Pronunciation:**
A	Alpha	Al fah
B	Bravo	Brah Voh
C	Charlie	Char Lee
D	Delta	Dell Tah
E	Echo	Eck Oh
F	Foxtrot	Foks Trot
G	Golf	Golf
H	Hotel	Hoh Tell (FAA, IMO, ITU) Ho Tell (ICAO)
I	India	In Dee Ah
J	Juliett	Jew Lee Ett
K	Kilo	Key Loh

Table 1. (continued)

L	Lima	Lee Mah
M	Mike	Mike
N	November	No Vem Ber
O	Oscar	Oss Car
P	Papa	Pah Pah
Q	Quebec	Keh Beck
R	Romeo	Row Me Oh
S	Sierra	See Air Ah (FAA) See Air Rah (ICAO, IMO, ITU)
T	Tango	Tang Go
U	Uniform	You Nee Form
V	Victor	Vik Tah
W	Whiskey	Wiss Key
X	X Ray	Ecks Ray
Y	Yankee	Yang Key
Z	Zulu	Zoo Loo

Note: The syllables with capital letters are to be stressed.

In addition to the letters, numbers in aviation English are pronounced based on the following recommendation (ICAO, 2001). Among these, a few numbers, such as 2, 3, 4, and 9, are distinctively pronounced to reduce ambiguity (see Table 2).

Table 2.

ICAO Numbers

Numeral or numeral element	**Pronunciation**
1	Wun
2	Too
3	Tree (not Three)
4	Fow-er (not Fow)
5	Fife
6	Six
7	Sev-en
8	Ait
9	Nin-er (not Nine)
0	Ze-ro

In aviation radiotelephony, whole hundreds and thousands are pronounced as each digit in the number of hundreds or thousands followed by the word “hundred” or “thousand.” Combinations of thousands and whole hundreds are pronounced with each digit in the number of thousands followed by the number of hundreds. All numbers except whole and combinations of hundreds and thousands are transmitted by pronouncing each digit respectively. Numbers with a decimal point are transmitted one by one whether the number is a whole hundred or whole thousand. When communicating time, each digit of the hour and minute are pronounced one at a time (Dejkunjorn, 2005). Examples of how to transmit numbers, decimal points, and time are provided in Table 3.

Table 3.

Examples of Numbers, Decimal Points, and Time

Number	**Transmitted as**
30	Three Zero
81	Eight One
512	Five One Two
800	Eight Hundred
3000	Three Thousand
5800	Five Thousand Eight Hundred
21000	Two One Thousand
45600	Four Five Thousand Six Hundred
68912	Six Eight Nine One Two
100.3	One Zero Decimal Three
2000.5	Two Zero Zero Zero Decimal Five
38243.8	Three Eight Two Four Three Decimal Eight
0800 (8:00 A.M.)	Zero Eight Zero Zero
15:35 (3:35 P.M.)	One Five Three Five

As in the prior explained cases of numbers, decimal points, and time delineations, the words and phrases used in aviation English are also standardized to avoid ambiguity. The nine basic principles of aviation phraseology are provided below:

1) The English language should be the basis for the development of the requisite phraseologies. Words with Latin roots should be given preference in developing the phraseologies;
2) Words and phrases should be selected in such a way as to ensure optimum transmissibility over radiotelephone channels and should be incapable of misinterpretation;
3) Words and phrases should be avoided which will be liable to differences of pronunciation likely to cause misunderstanding;
4) Spoken Q code groups, which by their common usage have already become part of aviation terminology, may be used where they provide a preference alternative to a long or complex phrase, e.g., QEE, QFF, QNE, QNH, QTEZ (QFE: Atmospheric pressure at aerodrome elevation or at runway threshold; QFF: Barometric pressure at a place; QNE: Landing altimeter reading when subscale set 1013 hectopascals; QNH: Altimeter subscale setting to obtain elevation when on the ground; and QTE: True bearing);
5) Phrases already in general use that have proved, by experience, to be phonetically suitable irrespective of the language from which they were derived should not be arbitrarily changed;
6) New phraseologies developed during the study should be clear, unambiguous, and, where practicable, concise. However, clarity should not be sacrificed in the interest of brevity;
7) Phrases should be developed on the principle that they present a thought expressed in a live language. However, the grammatical construction should be as simple as possible;
8) Positive and negative instructions or advice should be clearly differentiated;
9) Where practicable, words containing sounds or syllabic constructions traditionally

difficult to pronounce by non-English-speaking personnel should be avoided (ICAO, 2001, pp. ATT B-1-ATT B-2)

The following words and phrases introduced in Table 4 are commonly considered to be appropriate in radiotelephony communications and carry the meanings displayed below.

Table 4.

Examples of Aviation English Phraseology (ICAO, 2007b)

Phrase	**Meaning**
Acknowledge	"Let me know that you have received and understand this message."
Cleared	"Authorized to proceed under the conditions specified."
How do you read	"What is the readability of my transmission?"
Negative	"No" or "Permission not granted" or "That is not correct" or "not capable".
Roger	"I have received all of your last transmission."
Say again	"Repeat all, or the following part, of your last transmission."
Standby	"Wait and I will call you."
Wilco	"I understand your message and will comply with it."

The characteristics of aviation English communication between a pilot and an air traffic controller can be identified in the taxi instruction procedures presented in Table 5. In the given example, showing aviation English communication between a pilot and a controller, it is the pilot who initiates the transmission in order to request data for taxi information by calling the call-sign of the ATC tower (TWR). Both the pilot and the controller call the other party's call-sign as a vocative (addressing term) first when they initiate transmission. To clarify the received and transmitted information, the terms "readback" and "hearback" are frequently used by the pilot and the controller. In the example, there are three verbal steps involved in the entire taxi instruction procedure. The pilot is supposed to repeat the primary information transmitted by the controller, which is called the "readback." In step 5, the controller actively listens to the pilot's

readback to confirm that the pilot received transmitted data or instructions correctly, a verification which is called “hearback.”

Table 5.

Requesting Taxi Instruction: Pilot-Air Traffic Controller Communication Procedure Example (ICAO, 2007b)

Pilot	**Air Traffic Controller**	**Example 1** (Taxi Instructions)
1. Call control (TWR) - Name of ATC TWR - Aircraft call sign - Request information		- “Georgetown Ground” - “Fastair 345” - “Request Taxi information”
	2. Controller replies - Aircraft call sign - Runway direction - Further instruction	- “Fastair 345” - “Taxi to holding point runway 27 Departure runway 32” - “Give way to B747 passing left to right QNH 1019”
3. Pilot replies - readback - aircraft call sign		- “Holding point Runway 27 QNH 1019 giving way to B747” - “Fastair 345”
	4. Controller actively listens to check if the pilot’s repetition of the instructions is correct (hearback)	

As seen in the examples in Tables 1 through 5, clarity and brevity are emphasized in radiotelephony to prevent ambiguity and miscommunication by repeating readback and hearback of crucial information, including runway direction, altimeter, and instruction following, and by adopting concise aviation English phraseology.

Army Aviation English in Korean Contexts

Korean army aviation English, the target context of the dissertation study, also complies

with the aviation regulations issued by Federal Aviation Administration (FAA) and ICAO. Korean army pilots and controllers have been trained to have a firm command of the aviation phraseologies based on the international standards, so they can efficiently carry out joint military operations with U.S. Armed Forces in Korea. As the use of aviation English is supposed to remain the same regardless of the aircraft types, such as fixed-wing aircraft (e.g., a jet fighter), rotary-wing aircraft (e.g., a helicopter), glider, and hot-air balloon, pilots, and controllers in Korean Army Aviation have also adopted words and phrases recommended by FAA and ICAO. However, a unique characteristic of the target context is that all aircraft currently held by Korean Army Aviation are helicopters. Due to the features of rotary-wing aircrafts, which do not need a long runway to take off and land, mainly fly at low altitudes, and rely far more on visual flight rules (VFR) as opposed to instrument flight rules (IFR), the use of aviation English tends to add technical terms according to the qualities of the aircrafts and military operation types based on the ICAO-enacted aviation words and phrases (Jang, 2001).

Construct Definition of Aviation English

Language ability is an abstract concept; therefore, it is necessary to define ability in explicit terms for language assessment. Such defined ability, inferred from a meaningful interpretation of observed behavior, is called a construct (Bachman & Palmer, 2010; Chapelle, 1998). There are three theoretical perspectives in construct definition – trait, behavior, and interactionalist (Chapelle, 1998). A trait definition views a construct in terms of knowledge and fundamental processes of the test taker. A behaviorist definition defines a construct with reference to the contextual features wherein test performance is observed. Lastly, an interactionalist definition includes strategic competence which mediates both trait and contextual features.

The construct of interest is aviation English ability. Aviation English consists of two types of language: standardized phraseology and plain language. Standardized phraseology is defined as the language of routine air-ground communication consisted with brief, concise, and accurate transmission of specific information related to the flight (Emery, 2014). While, plain language/English, which is defined as "the spontaneous, creative and non-coded use of a given natural language" (ICAO, 2010, p. x), and shall be used only when standardized phraseology cannot serve an intended transmission" (ICAO, 2010, Appendix A). Pilots and controllers are expected to use standardized phraseology to reduce the risk of miscommunication; however, when it comes to abnormal or emergency situations, despite the existence of alternative standardized phraseology, they tend to utilize plain English (Kim & Elder, 2009; Howard, 2008). As highlighted recent studies (Douglas, 2014; Kim & Elder, 2009), plain English in the context of aviation communication is neither literally plain nor colloquial / general language, yet it is highly context-specific and purpose-driven, just not in the standardized form.

The construct to be assessed in this dissertation study is air traffic controllers' aviation English ability in Korean Army Aviation contexts. The target construct is based on the interactionalist definition (Chapelle, 1998) in which a test taker's performance on the test is taken to indicate an underlying trait characteristic of that test taker. At the same time, the performance is also taken to indicate the influence of the assessment task or situation context where the performance takes place. In this case, the construct is the ability of aviation English communication in the target language context of Korean army aviation, whose meaning includes a language ability trait whose scope is delimited by the army aviation context. Performance on the aviation English test is an indicator of aviation English ability produced in context. The test task performance is treated as a sample of performance in the ATC tower in the Army Aviation.

Moreover, an interactionalist construct definition involves not only trait and context, but also strategic competence [i.e., skills necessary to put language knowledge into use (Bachman & Palmer, 1996)] whereby the test taker utilizes her or his language knowledge (e.g., acceptable aviation English) in a given assessment context (e.g., Korean Army Aviation English) through strategic competence. Therefore, the construct definition for aviation English ability based on the interactionalist perspective would include the knowledge of aviation English communication in the context of the Korean military ATC situation with the strategic competence to direct and assess its use. The test scores, therefore, are intended to be indicators of this construct.

Aviation English Assessment. It was only a decade ago that the International Civil Aviation Organization (ICAO) published a set of Language Proficiency Requirements and a Proficiency Rating Scale. Air traffic controllers and pilots were expected to have a ICAO endorsed certificate to prove their English proficiency for international aeronautical communication (ICAO, 2004). Despite some testing organizations' efforts to develop and implement tests by the deadline, however, a new deadline of March 2011 had to be re-established. It was because many of the aviation English assessment procedures were found not to satisfy professional standards for language tests, implementation of the aviation English assessment policy was inadequate and much closer monitoring on the test quality and procedures is needed (Alderson, 2009).

For aviation English proficiency assessment, ICAO does not directly administer aviation English tests, but provides a general set of guidelines of aviation language proficiency requirement (e.g., LPRs). This open interpretation of the guidelines actually promotes the endeavors of private corporations, government institutions, or educational testing companies to develop a variety of aviation English tests (Alderson, 2010). Unfortunately, the access to recent

aviation English tests or related research studies seems to be limited. Like other ESP test development projects, much of the research on aviation English test development appears to be localized, and on-site ESP / LSP research is either unpublished, published internally, or in a language other than English (Barbieri, 2015). Currently, only two tests are certified by ICAO: the English Language Proficiency for Aeronautical Communication (ELPAC) developed by European Organization for the Safety of Air Navigation (EUROCONTROL) and the RMIT English Language Test for Aviation (RELTA) developed by the Royal Melbourne Institute of Technology (RMIT).

Though not certified by ICAO yet, one of the leading aviation English tests, Pearson's Versant Aviation English Test (VAET) adopts automated speech recognition (ASR) technology in a computerized testing platform for fully automated test administration and scoring (Van Moere et al., 2011). There are numerous empirical and theoretical research studies on the VAET, such as evaluation of test usefulness (Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2008); reliability of computer-generated scores (Downey, Suzuki, & Van Moere, 2010); test construction and scoring model development (Van Moere, 2010); and issues on human vs. machine rating and assessment of spoken interaction (Van Moere et al., 2011).

Considering the reputation of Pearson's VAET, which was developed under a co-operative research and development agreement with the Federal Aviation Administration (FAA) of the United States, the researcher observed that Professor Dan Douglas, one of the researcher's dissertation committee members and prominent scholar in language testing, actually passed the VAET test without any civil or military ATC training or service experience, while a non-native English speaking ATC Master Sergeant (MSgt) with 17 years ATC experience of the current dissertation research context failed the test. This anecdote is introduced not to find fault with the

VAET, but to highlight the importance of the two aspects of language testing authenticity: situational and interactional (Douglas, 2000). As Korean Army Aviation operates helicopters (rotary-wing aircrafts), the VAET, which mainly focuses on civil fixed-wing aircrafts, could have been be too challenging for the MSgt. In other words, the VAET may focus on language ability superficially in the context of aviation English and may not measure what he actually could perform in an authentic TLU situation.

Douglas (2000) highlights that it is not enough merely to give test takers topics relevant to the field they are studying or working in. The test materials must engage test takers in a task in which both language ability and knowledge of the field interact with the test content in an authentic way similar to the TLU situation. The importance of authenticity motivated the researcher to revisit the TLU situation and led the researcher to adopt a task-based approach (e.g., task-based needs analysis, task-based language assessment) for the development of more authentic and localized aviation English test tasks which will be explained more in the following section.

Issues identified from the recent studies (e.g., Alderson, 2009; Van Moere et al., 2011) concern the ICAO Language Proficiency Rating Scale which consists of six levels of skills in six areas of language use: pronunciation, structure, vocabulary, fluency, comprehension, and interaction. ICAO set Operational Level 4 for all of the six criteria on the 6-point scale (i.e., Level 6: Expert; Level 5: Extended; Level 4: Operational; Level 3: Pre-operational; Level 2: Elementary; Level 1: Pre-elementary) as the minimum standards for language proficiency for pilots and air traffic controllers. The definition of Operational Level 4 in each criteria are as follows (ICAO, 2007a):

Pronunciation (Assumes a dialect and/or accent intelligible to the aeronautical community) Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation, but only sometimes interfere with ease of understanding.

Structure (Relevant grammatical structures and sentence patterns are determined by language functions appropriate to the task.) Basic grammatical structures and sentence patterns are used creatively and are usually well controlled. Errors may occur, particularly in unusual or unexpected circumstances, but rarely interfere with meaning.

Vocabulary

Vocabulary range and accuracy are usually sufficient to communicate effectively on common, concrete, and work-related topics. Can often paraphrase successfully when lacking vocabulary in unusual or unexpected circumstances.

Fluency

Produces stretches of language at an appropriate tempo. There may be occasional loss of fluency on transition from rehearsed or formulaic speech to spontaneous interaction, but this does not prevent effective communication. Can make limited use of discourse markers or connectors. Fillers are not distracting.

Comprehension

Comprehension is mostly accurate on common, concrete, and work related-topics when the accent or variety used is sufficiently intelligible for an international community of users. When the speaker is confronted with a linguistic or situational complication or an

unexpected turn of events, comprehension may be slower or require clarification strategies.

Interactions

Responses are usually immediate, appropriate, and informative. Initiates and maintains exchanges even when dealing with an unexpected turn of events. Deals adequately with apparent misunderstandings by checking, confirming, or clarifying.

The set of language proficiency requirements (LPRs) has served as the construction definition of aviation English proficiency and the basis for test construction and for the rating of language proficiency. Current ICAO Language Proficiency Rating Scale (see Appendix C) may serve its intended purpose (test use) of measuring six aviation English constructs. However, test results according to the ICAO LPRs may not be the best indicator of aviation English ability of the current dissertation context, as test users of the current dissertation context of Korean Army Aviation expect the test results can specifically indicate what kinds of target tasks the test takers were able to perform successfully or vice versa for retraining of novice controllers. This intended purpose of the test and identified needs of the test users guided the researcher to design and develop an authentic task-based assessment as well as a task-centered rating rubric so that test results could provide diagnostic feedback to the test users and test takers.

It is important to note that the construct of aviation English ability in an actual TLU situation is far richer than in an aviation English test setting (Douglas, 2000). Though it may happen to a greater or lesser extent, air traffic controllers utilize cognitive and/or metacognitive strategies to perform target tasks in a real ATC environment or in a testing environments. Cohen (2014) highlighted in his book chapter *"Strategy use in testing situation"* that test-taking

strategies should be also taken into account both in designing and validating tests. Yet, reflecting the researcher's hands-on experience of air traffic control, conventional practice of aviation English assessment does not seem to promote test takers' strategy use.

Current aviation English tests (e.g., Versant Aviation English Test (VAET) by Pearson, Test of English Language Proficiency for Aeronautical Communication (ELPAC) by EUROCONTROL, Test of English Language Level for Controllers and Pilots (TELLCAP) by TELLCAP, and RMIT English Language Test for Aviation (RELTA) by RMIT University) are promoted with a computer-assisted testing platform. However, for example, VAET test merely adopts a computer screen as a digital version of the test sheet, displaying only a few images in the test items, and actually failing to facilitate authentic testing environments. The more the testing environment is authentic, the more strategic competence the examinees may use in the task performance (Douglas, 2001).

As Douglas highlights, if we want to assess how well test takers can use language for specific purposes, we require a measure which can consider both their language knowledge and background knowledge, and their use of strategic competence related to the target language use situation. By simulating authentic target tasks in a 3D virtual environment, the virtual interactive tasks for aviation English assessment (VITAEA) of the current dissertation study aim to assess the level of task accomplishment and six constructs under the ICAO's language proficiency requirements using language-centered and task-centered rating rubrics. Although the distinction between standardized phraseology and plain language is not the primary concern of the study, authentic target task performance in a simulated virtual environment facilitates test takers' use of plain English as well as standardized phraseology.

Task-Based Performance Assessment

Performance assessment can be defined as any test designed to elicit performance of specific language behaviors in the real world or a simulation of a real-life activity (Bachman, 2002; Norris, Brown, Hudson, & Yoshioka, 1998; Norris, Hudson, & Bonk, 2002). Skehan (1998) reasons that predominant approaches to language testing (e.g., Bachman's 1990, and Bachman and Palmer's 1996, models of strategic competence) have over-emphasized the search for an underlying structure of abilities, which determine test takers' performances. As an alternative, he suggests to investigate performance and processing considering test performance as a sample of performance that can be expected in similar contexts.

Performance tests are designed to elicit performance of simulated authentic language use, which can predict future real-life performances. They are typically rated based on the study of what is valued in a real-world context. Advocates of performance assessments highlight the advantages of such assessments, such as how the tests provide more valid estimates of test takers' true language performance to respond to real-life language tasks; and predictions of test takers' future performances in real-life language situations. Moreover, well-designed performance assessments can also be used to counteract the negative washback effect of standardized assessment, and provide strong positive washback effects (Brown & Hudson, 1998). In general, washback is perceived as being either negative or positive. Negative (harmful) washback could take place when the test content or format is based on a narrow definition of language ability and, consequently, restricts the context of teaching and learning. On the other hand, positive (beneficial) washback results in encouraging 'good' teaching and learning practices (Davies, 1999).

Task-based performance assessment fits within the broader definition of performance assessment. Especially, in task-based performance assessment, success in performing the tasks is a central issue (Brown, 2004). In other words, task-based performance assessment does not simply employ the real-world task as a means to elicit specific components of the target language system which are then evaluated or measured; on the contrary, the construct of interest in task-based performance assessment is success in performance of the task itself (Long & Norris, 2001). In this regard, authenticity of the target tasks is an essential quality in all performance assessment. It is expected that the closer the relationship between target tasks and real-life situations, the more accurate the generalization of test scores to non-testing situations will be.

As an initial phase of task-based performance assessment development, the current dissertation study adopted a task-based needs analysis which employs tasks as the unit of analysis and a task-based performance assessment design. Two critical advantages of a task-based needs analysis: (a) a task-based needs analysis identifies the target language use in real-world situations using the dynamic qualities of the target discourse; and (b) the results of a task-based needs analysis can be readily used as authentic input for the task-based language assessment as well as for task-based lessons or course design (Long & Norris, 2000).

To help the test task developers explore ways to differentiate and sequence assessment tasks, task difficulty can be determined according to their difficulty levels (Norris et al., 1998; Skehan, 1996). The three variables of central interest in the task difficulty index are (1) code complexity, (2) cognitive complexity, and (3) communicative demand. The code complexity of a given task addresses the kind of language and information that is involved in successful task performance, and focuses on the (a) range and (b) number of input sources. Range indicates the extent to which the code that is inherent in the language of a given task represents a greater or

lesser degree of spread. Number of input sources represents whether or not the examinee must decode multiple sources of information input. The cognitive complexity of a given task turns on the amount and kind of information processing that a test taker must engage in to successfully perform the task. The estimated cognitive complexity of the tasks is based on the two variables: (a) input/output organization and (b) input availability. Input/output organization specifies the extent to which information must be significantly organized for successful task completion. Input availability describes whether or not a test taker is required in some significant way to search for the information upon which a task performance is to be based. The communicative demand of a given task is established by the type and the number of moderator variables of communicative language activity with the two subsets of (a) mode and (b) response level. Mode indicates whether a given task is interpreted to have a productive component for successful task accomplishment. Response level denotes the extent to which a test taker must interact with input in a real time sense. The difficulty level ranges from zero (least difficult) to six (most difficult) based on the summed up numeric value of the six variables (code complexity, cognitive complexity, input/output, input availability, mode, and response level) with zero (generally lower estimated difficulty) or one (generally higher estimated difficulty).

Prior to the actual test development in the current dissertation study, a task-based needs analysis (Long & Norris, 2000; Long, 2005; Norris et al., 1998) was adopted to explore the target aviation English tasks in the TLU situation with perceived importance and analyze difficulty of the tasks, as well as to develop the task-centered rating criteria under the task-based performance assessment framework. In task-based performance assessment, the construct of interest is the performance of the task itself following six basic components.

Table 6

Six basic components of task-based language assessment (Norris, 2002)

1. **The intended use**(s) for task-based assessment within the language program must be specified, minimally addressing the following four issues:
- who uses information from the assessment?
- what information is the assessment supposed to provide?
- what are the purposes for the assessment?
- who or what is affected?
- what are the consequences of the assessment?

2. **Target tasks or task-types** emerging from the needs analysis are analyzed and classified according to a variety of task features. Analysis is undertaken in order to understand exactly what real-world conditions are associated with target tasks and should therefore be replicated under assessment conditions.

3. Based on information from the analysis of task features, **test and item specifications** are developed. Specifications delineate the formats tests should take, procedures involved, tasks or task-types to be sampled, format for test tasks (items), and how performance on the task-based test should be evaluated.

4. Carrying out **identification and specification of rating criteria**, which form the basis for interpretations of examinee performance and task accomplishment. Real-world criterial elements (aspects of task performance that will be evaluated) and levels (descriptions of what success looks like on these aspects of task performance) should be identified within initial needs analysis, with a view toward providing students and teachers with clear learning objectives.
5. **Task items, test instruments and procedures and rating criteria need to be evaluated** (involving pilot-testing and revision) according to their efficiency, appropriacy, and effectiveness with respect to the intended assessment uses.

6. Finally, task-based language assessment should **incorporate procedures for systematic and ongoing validation of its intended use** within the language program. Validation should minimally consider: to what extent test instruments and procedures are providing appropriate, trustworthy and useful information; to what extent particular uses for the assessment are warranted, and to what extent the consequences of assessment use can be justified, given the impact on students, teachers, language programs and any other relevant stakeholders in the assessment process.

Despite the potential advantages of task-based performance assessment with its six components, the researcher noticed vagueness in the approach to meet the six components, especially in task specification (Bachman, 2002) and procedures for validation which could inevitably lead to test validity and reliability issues. Literature (Brown & Hudson, 1998; Norris

et al., 1998) also shows that validity may be problematic owing to inadequate content coverage; lack of construct generalizability; the sensitivity of performance assessments to test method, task type, and scoring criteria; construct underrepresentation; and construct-irrelevant variance. Reliability also could be an issue due to inconsistent rating, limited numbers of observations, and rater subjectivity in the scoring process, to name a few issues. Furthermore, another challenge of performance assessments concerns logistical issues such as designing authentic task environments, collecting and storing audio or video data of test taker performance, providing test taker training, and ensuring test security

As a more reasonable possible solution to the conundrum of complexity in performance assessment, the evidence-centered design (ECD) model proposed by Mislevy, Steinberg, and Almond (2002) has been recommended in several foundational papers (Brown, 2004; Hines, 2010) and is introduced in Chapter 3. Assessing complex interactions in the VIAET requires a comprehensive framework for making valid inferences about virtual assessment interaction. The ECD framework (Mislevy, Steinberg, & Almond, 2003) provides a formal, multi-layered approach to designing assessments as arguments (Mislevy & Haertel, 2006; Mislevy, et al., 2003). The ECD framework helps to make explicit how rich assessment data in the VIAET are to be established through iterative cycles of analysis, design, development, implementation, and evaluation instructional design decisions. Using this approach, an evidentiary assessment argument can be formed that connects evidence and supporting rationales. Such an approach/perspective would essentially provide a framework for developing target tasks that elicit evidence to directly support the claims the researcher wants to make about what test takers can do (Mislevy, et al., 2003).

As another way to solve the troubling issues of practicality and logistics, and to develop more authentic task environments, the researcher adopts virtual worlds (VWs) as a test platform for aviation English assessment. As the goal of aviation English assessment is providing valid inferences based on the observation of test takers' aviation English performances (see Figure 1), task-based language assessments developed in VWs are able to offer various forms of evidence and methods in one assessment. Furthermore, language assessment based on authentic tasks and virtual world technology will potentially be more practical, cost effective, reliable, and valid than conventional paper-and-pencil tests.

Virtual Worlds

The discussion in this literature review so far has focused on the theoretical bases for aviation English under the umbrella of language for specific purposes as an assessment domain, and task-based performance assessment as an assessment method. In this subsection, the researcher introduces virtual worlds (VWs) as a potential assessment environment for performing target test tasks as well as supple compensating the shortcomings of task-based performance assessment.

Definition and Characteristics of Virtual Worlds

Virtual worlds (VWs) are simulated environments where avatars represent users who interact and collaborate with each other, contributing to a sense of community belonging or social connectedness (Bell & Trueman, 2008; Peterson, 2010). Unfortunately, however, VWs have been mainly used for training or testing in science not in target language practice or social interaction with other avatars, but no known attempts seem to be made to incorporate VWs for

language assessment. Although VWs are not explicitly created for language education purposes, several language educators and researchers have made use of them as a platform to teach and learn a variety of disciplinary content, including languages (Silva, 2012). Currently, there are hundreds of VWs in which educators show interest. Table 7 summarizes some popular VW platforms and includes information about their costs, target audiences, themes, number of members, and installation method. As presented in Table 7, most VWs can be used for free except World of Warcaft (WoW), one of the most popular Massively Multiplayer Online Role-Play Gams (MMORPGs).

Table 7.

Overview of Popular VW Platforms

	Cost	Audience	Theme	Members	Installation
Active Worlds	Free	Unstated	Social	3+ million	Browser add-on
Club Penguin	Free	6-14	Penguins	200+ million	Browser-based
Habbo	Free	13+	Hotel	260+ million	Browser-based
Second Life	Free	13+	Social	36+ million	Download
World of Warcraft	$14.99/ month	Unstated	Role play, battle	12+ million	Download

All of the VWs, including the five platforms, share the following common characteristics (Sadler, 2012):

- *On-line 3D environment*. This may simulate the real world, a historical reconstruction, or something that exists only in fiction (e.g., a space city);
- *Avatars*. Avatars are the in-world representations of the real world people who control them;

- *Real-time interactivity.* VWs include the possibility of interacting with other avatars in the environment in real time (synchronous communication), and usually with a range of objects in that VW;
- *24-hour accessibility*. As opposed to interacting with a friend via a program like Skype™, where the potential for interaction only exists as long as the program is open and both parties are connected, a VW remains open and accessible 24 hours a day;
- *Persistence*. When a user logs out of a VW, their avatar, and the actions taken by that avatar, are not deleted;
- *Social space*. Although VWs may vary in look and theme, all are primarily social spaces that exist for the purpose of humans interacting via their avatars;
- *Numbers*. In most VWs, there are many players (sometimes in the hundreds of thousands) online in the world at the same time. Given the social nature of these spaces, underpopulated VWs tend to fade away quickly as users lose interest.

The popularity of VWs based on the strength of such attractive characteristics introduced above may be among the motivating factors drawing the attention of educators and students. Furthermore, there exist both theoretical underpinnings as well as research findings which confirm the advantages of VWs. One theoretical aspect underlying VWs is experiential learning theory, which defines learning as "the process whereby knowledge is created through the transformation of experience" (Kolb, 1984, p. 41). Additionally, project-based instructional activities in VWs have also been found to offer an adequate environment for such experiential learning cycles in higher education settings (Leifer, 1996). VWs provide users with highly interactive arenas for dynamic feedback, learner experimentation, real-time personalized task

selection, and exploration (Kalyuga, 2007). Other instructional benefits of VWs include promoting social interaction (Amichai-Hamburger & McKenna, 2006), creating community (Lamb, 2006), facilitating collaboration (Monahan, McArdle, & Bertolotto, 2008), lowering social anxiety (Barab, Thomas, Dodge, Carteaux, & Tuzan, 2005), and enhancing learner motivation and engagement (Dede, Clarke, Ketelhut, Nelson, & Bowman,2005).

Second Life as a Test Platform

Among the VWs shown in Table 6, Second Life (SL) has been one of the most widely explored platforms for users, educators, and researchers. Unlike most VWs, SL does not have a fixed purpose or theme other than social interaction in the virtual worlds. Instead, SL allows its users to make their own virtual land called a simulator (sim) and to design the land, avatars, and objects in the sim. One unique feature of SL is its internal economy and internal currency, the Linden dollar (L$). Linden dollars can be used to buy, sell, rent, or trade land or goods and services with other users. The "Linden" can be exchanged for U.S. dollars or other currencies on market-based currency exchanges (Second Life, 2014). Contrary to most VWs, which are aimed at engaging younger users, SL has no communication restrictions among different age groups. Thanks to a *landmark* function, which is similar to an Internet bookmark, experienced users, educators, and researchers in SL can link their own virtual space with the landmark, so they may conveniently invite other users. Based on the distinct characteristics of SL, it has a strong community of educators, numbering in the thousands, including teachers and researchers working with various languages (Sadler, 2012).

According to a recent survey given to 237 SL users (Sadler, 2011), more than 80% of SL participants have experienced a great deal of second language practicing with other users in SL

by attending a school in the online environment, listening to other users, reading texts (via chat functions), speaking in audio chat, and writing text chat. Sadler's survey study highlighted that SL can be an attractive environment for language learners to learn and practice a target language based on the following evidenced features: (1) users spend many hours in this world; (2) users are diverse in terms of their nationalities and language backgrounds; and (3) many users practice additional languages while in-world (Sadler, 2011).

Virtual Worlds for Language Assessments

The use of VWs could improve language assessment dramatically by validly measuring test takers' sophisticated intellectual and psychosocial performances by compensating the shortcomings of task-based performance assessment. First, VWs can improve face validity which refers to test takers' intuitive judgment about the test (Alderson, Clapham, & Wall, 1995) and which also corresponds to evaluation inference in the validation framework of the current dissertation research. The evaluation inference links the task performance observation with an observed score based on the warrant that the observed scores reflect test takers' targeted language abilities. Though frequently dismissed by testers as being unscientific and irrelevant (Stevenson, 1985), face validity can be important in determining its acceptability and reasonableness to those who will be tested and to those who use test results (Messick, 1989). Perceived relevance of a test was found to be indispensable for motivating test takers to demonstrate their full repertoire of skills (Anastasi, 1988). Simulating TLU situations in a virtual air traffic control tower and providing test takers with authentic performance tasks in the VW may also encourage test takers to fully demonstrate what they actually can do in the TLU situations.

Second, VWs can enhance construct validity, which corresponds to domain description inference in the validation framework of the current study. As discussed in the construct definition of aviation English, strategic competence serves as a mediator between test takers' internal traits of background and language knowledge and external context by controlling the interaction between them (Douglas, 2000). As Douglas emphasizes, it is test developers' a major responsibility for providing sufficient contextual information to enable the test takers to establish the context. Conventional aviation English tests merely use textual prompts on a piece of a paper or hypertext prompts and photographs on a computer screen. Such prompts definitely promote the use of strategic competence; however, strategic competence adopted in the conventional aviation English tests and virtual interactive tasks simulated in a virtual air traffic control tower cannot be identical. In other words, virtual interactive tasks in a virtual ATC control tower will surely provide enhanced strategic competence use over conventional aviation English tests. By approximating authentic TLU situations in VWs, test takers' use of strategic competence should be promoted to better predict future real-life performance.

The process of reasoning from test takers' task performance in an authentic virtual environment can be portrayed as the assessment triangle (Pellegrino, Chudowsky, & Glaser, 2001) shown Figure 1. The corners of the triangle represent the three essential elements underlying aviation English assessment for the current dissertation study. *Cognition* corner indicates the set of knowledge, skills, and competence needed for aviation English communication. *Observation* corner refers to target tasks selected for inferring cognition based on observed performance, the set of knowledge, skills, and competence. *Interpretation* corner means models for interpreting test takers' task performance collected from the observation.

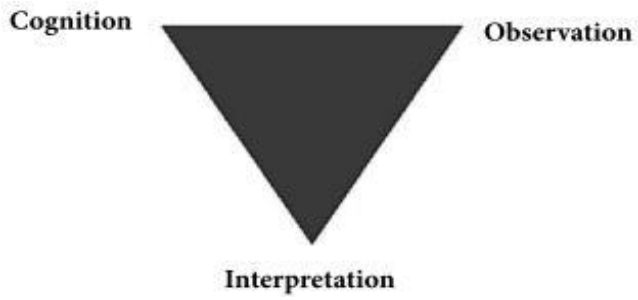

Figure 1. Assessment Triangle (Pellegrino, Chudowsky, & Glaser, 2001)

Despite the advancement of cutting edge technology, the *observation* corner of the triangle in the conventional army aviation English test has been primarily addressed with paper-and-pencil item-based tests. Such paper-and-pencil item-based tests, which force test takers to select among a few predetermined choices, fail to produce a wide scope of observations, but generate only a weak observation of whether they have acquired a sophisticated ability or skill involving advanced knowledge. Indeed, multiple-choice tests (e.g., conventional aviation English test in Korean Army Aviation) have been preferred by the test users in the military, as they are more cost effective, easier to rate, and have met psychometric criteria. Still, in spite of the limitations of performance-based assessments, advances in technology (e.g., VWs) can open up new possibilities and potentials for assessment.

The immersive interface and functions of VWs such as Second Life actually enable the emulation of complex situations with implicit clues, simulation of scientific instruments, virtual experimentation, simulated collaboration in a team, and adaptive responses to test takers' choices by capturing and recording responses in real-time in non-verbal, textual, and verbal modes (Dede, 2009). Moreover, it has been found that test taker performance on multiple-choice tests does not necessarily reflect the type of learning that we may observe via analysis of test takers' log file data as the users interact in VWs, conduct interviews or observations, and create summative essays (Clarke, 2006; Ketelhut, Dede, Clarke, Nelson, & Bowman, 2008).

In this regard, to establish the fairness, reliability, and validity of complex intellectual aviation English assessments, attempts should be made to facilitate and synchronously document observations of test taker performance, synchronously audio-recording and even video-taping every aspect of the performance by the aid of technical advancement. The type of observations and evidence of mastering knowledge and skills that authentic and performance-based assessment in VWs allows would be unparalleled.

Research Issues

Reviewing the concepts associated with the current dissertation study on LSP, aviation English, task-based performance assessment, and virtual world led to a good number of research issues with regards to the current dissertation study. One of the criticisms of LSP testing is about identifying representative samples of the behavioral domains of interest. To better identify the target tasks in the TLU situation, the researcher intentionally narrowed down the TLU situation of the current dissertation study to be army aviation English in the ATC tower context. However, identification of representative target task samples is a challenging job, and this issue will be more closely attended to through a task-based needs analysis and interviews with target domain experts.

Reviewing Pearson's VAET test, the researcher took an example of two test takers and pointed out the limitation of six aviation English construct-centered rating scale. In particular, it seems to be interesting how two different lenses of the rating rubric (language-centered and task-centered) could differentiate test takers' aviation English ability and which rubric could provide more useful or accurate inferences about the test users and test takers. To identify the relationship between the rating rubrics, semi-structured interviews and surveys with the test takers and test users needed to be conducted.

To compensate for the recognized limitations of performance-based assessments and to improve task authenticity, Second Life, a virtual world, was introduced as new possibilities and potentials for assessment (Clarke-Midura & Dede, 2010). The researcher become very curious about whether this endeavor of transforming the test environment into a virtual world could actually promote more cognitive and metacognitive strategies such as those that occur in real world task performance. To capture what specific strategies the test takers actually use is a difficult process to operationalize. However, the researcher will adopt a verbal protocol method to identify test takers' strategy use during their task performance. These issues will be incorporated as a part of the research questions and discussed further in the next chapter.

CHAPTER 3

TEST DEVELOPMENT AND VALIDATION

This chapter describes the development and validation of virtual interactive aviation. English tasks within the conceptual framework of an argument-based approach. The following sections describe Evidence-Centered Design (ECD) the theoretical underpinning in the test development. Then, the actual test development process is presented. The sections of test development are followed by an overview of the argument-based approach to validation which includes test interpretations, uses, and consequences in an interpretive argument. This chapter ends with a list of research questions for the current dissertation study, which were developed based on the backing required to support the interpretive argument to develop it into a validity argument.

Evidence-Centered Design for Test Development

During the past decade, researchers have made significant advances in multiple facets of assessment design. Current technological advances (e.g., virtual worlds) offer exciting opportunities to design assessments that are active and situated, measure complex student knowledge, and provide rich observations for student learning. Frameworks such as the Assessment Triangle (NRC, 2001) and ECD provide rigorous procedures for linking theories of learning and knowing to demonstrations of performance and to interpretation. ECD has proven useful as a design framework for a number of simulation-based assessments, especially in the science education field (Mislevy, Steinberg, & Almond, 2003; Mislevy & Haertel, 2006).

To provide a useful framework for test design and development in this study, a systematic assessment framework called ECD (Mislevy, Steinberg, & Almond, 2003) was used

to develop interactive performance assessments for measuring aviation English proficiency that could be reflective of situated, complex performances. “Evidentiary reasoning”, the heart of ECD, is an application of the concept of argument to test design and development. The chain of reasoning in ECD concerns everything from what test takers say and do in assessment to inferences about what they know or can do (Mislevy et al., 2003). Test items, tasks, and test specifications were developed based on such systematic and evidentiary reasoning. ECD is/was able to function not only as a test design and construction blueprint, but also as *a priori* validity evidence. In addition, the adoption of the ECD framework could also add support to the shortcomings of the TBLA framework by providing rigorous procedures for linking theories of learning and knowing to demonstrations and interpretation, thereby addressing the need to fill exposed reliability and validity gaps in TBLA.

Consequently, adopting ECD in test design and development can be beneficial to test designers who need to both justify the appropriateness of test design and development for a proposed use and investigate evidence about the validity of test use. In the current dissertation, the ECD framework as a methodology that comprises practices for creation and on-going development of aviation English assessment is adopted from Mislevy et al. (2003) (see Table 8).

Table 8

Mislevy’s Four-Stage Evidence-Centered Design Process for the Virtual Interactive Tasks for Aviation English Assessment (Mislevy et al., 2003)

Stage	Process	Component	Definition of component
1. Domain Analysis	Preliminary synthesis of what is known about what is to be assessed	No specific components specified although useful categories enumerated	NA

Table 8. (continued)

2. Domain Modeling	Incorporation of information from stage one into three components; sketch of potential variables and substantive relationships	Proficiency paradigm Evidence paradigm Task paradigm	Substantive construct expressed as claims Observations required to support claims Types of situations that provide opportunities for test takers to show evidence of their proficiencies
3. Construction of Conceptual Assessment Framework	Development of a final blueprint; provide technical detail required for implementation including statistical models, rubrics, specifications and operational requirements	Student model	Statistical characterization of the abilities to be assessed
		Evidence model	1. Rules for scoring test tasks 2. Rules for updating variables in the student model
		Task model	Detailed description of assessment tasks
(Assessment Implementation)		Presentation model	Specification of how the assessment elements will look during testing
		Assembly model	Specification of the mix of tasks on a test for a particular student
4. Deployment of Operational Assessment (Assessment Delivery)	Construction of the operational delivery system	Presentation	Presentation, interaction and response capture
		Response scoring	Evaluation of response; task-level scoring
		Summary scoring	Computation of test score; test-level scoring
		Activity selection	Determine what to do next

The first stage of test development, domain analysis, focuses on the preliminary synthesis of what is known about aviation English. This stage helps test developers understand the kinds of problems and situations non-native English speaking air traffic controllers deal with, ATC knowledge and skills they draw upon, the representational forms of aviation English they use, and the characteristics of good ATC task performance. Salthouse (1991) notes that for performance-based assessments, designers must understand the kinds of situations that pose recurring problems. Actual air traffic controllers' patterns of thinking and performing in ATC situations will be at the heart of the assessment, in forms and at levels that match the purpose(s) of the assessment purposes; hence, expert practitioners' and researchers' insights about the target

language and target context are invaluable. Through a task-based needs analysis, the characteristics of test takers, target contexts, target tasks, and situations are identified.

In the domain modeling stage, findings from the domain analysis are organized as an assessment argument (Mislevy, 2003, 2006). The assessment argument describes the relationship among target knowledge and the skills that test takers are expected to demonstrate, and those tasks and situations that evoke such knowledge. An extended Toulmin argument diagram for an assessment argument is provided in Figure 2.

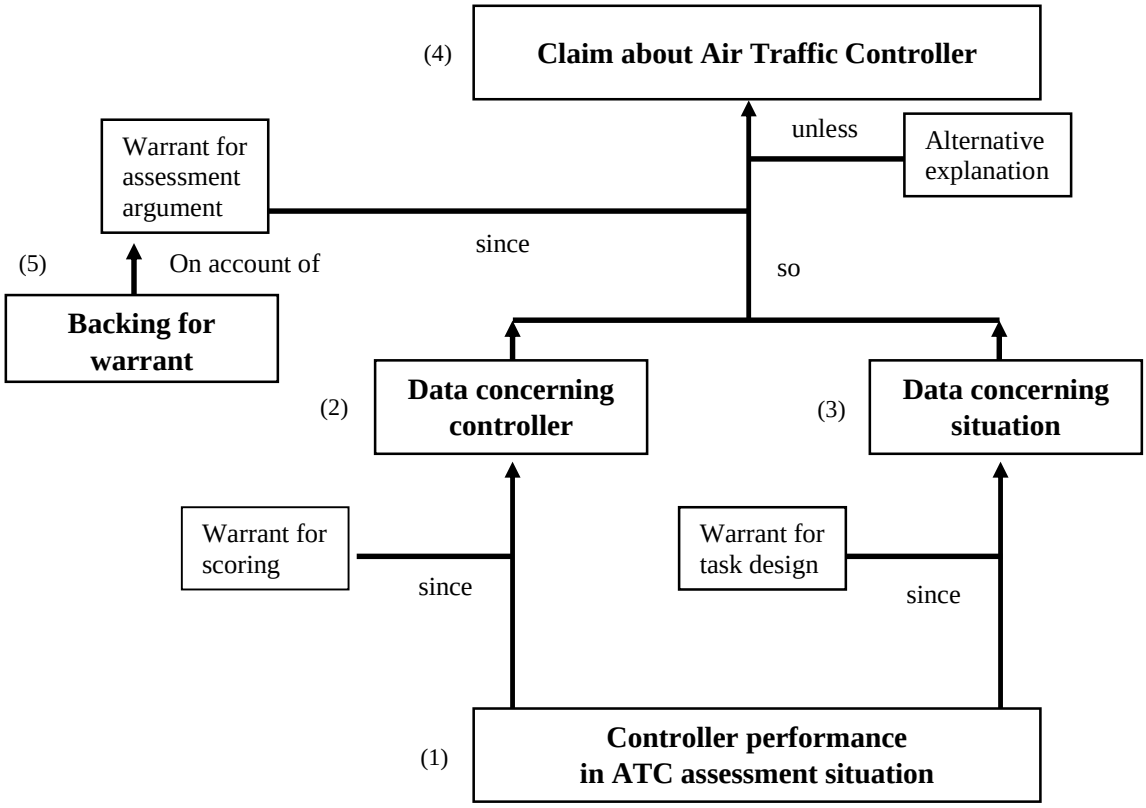

Figure 2. An Extended Toulmin Argument Diagram for Assessment Argument

Figure 2 is intended to illustrate that a claim about a test taker (air traffic controller) is justified by assessment data through a warrant. At the bottom, (1) controller performance in the ATC assessment situation indicates what a controller (test taker) says, does, or how the person interacts with a simulated pilot throughout the task completion. The combined data concerning

(2) the test taker (controller) and (3) situation are crucial for evaluation of performance. First, the situation needs to include features that are needed in order to evoke the target language. Second, the expected target task performance should be practiced so that the knowledge and skills used in the task performance do not go beyond the target knowledge and skill. Then, (4) the claim about a test taker can be justified by assessment data through

Based on the findings from the domain analysis, domain modeling incorporates the following three components in the development of the assessment (Hines, 2010):

1. Proficiency Paradigm – What substantive claims will be made about the test takers' abilities or competences?
2. Evidence Paradigm – What observable features in test takers' performances would provide data to support these claims?
3. Task Paradigm – What kinds of tasks provide an opportunity for test takers to demonstrate evidence of their proficiencies?

Designing assessment arguments for complex knowledge and skills in aviation English is challenging; however, one of the advantages of ECD is the use of assessment design patterns (Mislevy, et al., 2003). Assessment design patterns help test developers sketch out a design space for certain challenging-to-assess aspects of knowledge and skill by (1) building around the elements of an assessment argument; (2) incorporating experience from research and previous experience; and (3) providing flexible forms of assessment for a range of purposes (Mislevy, 2013). Table 9 is an abridged design pattern which describes the design of aviation English tasks and was used to guide creation of specific aviation English tasks in a virtual environment and determining what evidence to seek in more complex tasks.

Table 9.

A Design Pattern to Support the Tasks for Aviation English Assessment

Attribute	Value(s)
Name	Aviation English communication in a Korean Army Aviation context
Overview	Build on task-based approach, using a virtual environment called Second Life, this design pattern concerns identifying direct evidence about aspects of the aviation English communication abilities in a given context.
Central claims	Aviation English communication abilities in the Army Aviation context to demonstrate aviation English ability to speak and understand the language used for radiotelephony communications.
Additional knowledge that may be an issue	Knowledge of computer system components and functions; familiarity with Second Life, a virtual environment; limitation of the virtual environment compared to reality.
Characteristic features	Procedures for task performance by interacting with a simulated pilot and environments.
Variable task features	Complexity of tasks. Scope: Interaction with accessible information (e.g., weather, flight plan); interaction with simulated pilots for departure, landing, and transition. Setting: Interactive simulation, non-interactive simulation. Type of fault: single, multiple, constant, or intermittent. Kind / degree of support: warm-up session before the assessment, synchronous advice from task developer in the same virtual environment.
Potential performances and work products	Trace & time stamps of actions and speaking; video of actions in Second Life; stimulated recall (verbal protocol); explanations of test results; completed representations.
Potential features of performance to evaluate	*Regarding the performance*: Systematic vs. erroneous sequences of ATC; serial elimination vs. redundant or irrelevant ATC communication; efficiency of ATC communication. *Metacognitive:* interpretation of situation; selected strategies in problem solving.
Selected references	ICAO. (2007b). Manual of radiotelephony: International Civil Aviation Organization. ICAO. (2009). Guidelines for Aviation English Training Programs. Kim, H. (2012). Exploring the Construct of Aviation Communication: A Critique of the ICAO Language Proficiency Policy. Van Moere, A., Suzuki, M., Downey, R., & Cheng, J. (2011). Implementing ICAO language proficiency requirements in the Versant Aviation English Test. Kim, H., & Elder, C. (2011). Understanding aviation English as a lingua franca: Perceptions of Korean aviation personnel.

The Conceptual Assessment Framework (CAF) as the third stage of ECD emphasizes the combination of domain information, including goals, constraints, and logistics, to create an assessment blueprint. The CAF is composed of three central models: (1) student model; (2) task model; and (3) evidence model. The student model includes statistical characterization of test takers, which means their number and character depend on the purpose of the assessment. For example, a single-student model can be used to characterize a test taker's overall proficiency in a domain of tasks, while a multidimensional student model can be used to identify proficiency patterns in complex performances and/or to provide more specific feedback. The test development in this dissertation study adopts a multidimensional student model, as the test takers are expected to produce complex performance in the targeted tasks. Additionally, test users expect to provide more detailed diagnostic feedback to test takers (Mislevy, 2011).

A task model provides specific descriptions of test tasks environments, which is, in many aspects, similar to task specification. As the aviation English assessment in this dissertation study is implemented in a virtual environment, detailed information is required to specify a task. For this study, a task is specified with its initial status, transition rules, and the form(s) in which a test taker's performances is captured. An evidence model connects the student model and the task model with two subcomponents: (1) evaluation and (2) measurement. The evaluation component involves the rules for task scoring. For this aviation English assessment, a language-centered rating rubric (based on the ICAO's LPRs rating scale) and a task-centered rating rubric (derived from the findings based on the empirical research on the TLU situations) are adopted to provide detailed diagnostic feedback as well as the level of specific task accomplishment. For linear tests, the Assembly Model corresponds to a test blueprint which provides detailed requirements that must be satisfied for each test form (e.g., number of tasks, content to be covered, the order of the

tasks). Lastly, the Presentation Model concerns the formatting specifications for a test. As the fourth stage of ECD, the deployment of operational assessment is made up of a four-process delivery system. This assessment delivery stage includes: (1) presentation of information, interaction between a test taker and the information, and the capture of a test taker's response; (2) scoring of the response; (3) summarization of the scores throughout the responses; and (4) decisions about possible steps in the future (Hines, 2010).

Task Design Analysis

To provide clear articulation of the connection between complex elements of ECD and actual test design, a task design analysis (TDA) (Hines, 2010) was conducted including a task-based needs analysis study. This section delineates the components and stages of TDA followed by findings from the task-based needs analysis study which aimed to correspond to each stage of the TDA. Based on the results from the task-based needs analysis, the outcomes from the six-step TDA for the aviation English assessment are provided.

Regarding the structure of TDA, this six-step approach in TDA primarily focuses on the first two stages of ECD, domain analysis and domain modeling (see Table 10).

Table 10.

Steps Carried Out in Task Design Analysis Guided by Aspects of Evidence-Centered Design (Hines, 2010)

Step in task design analysis	Component from evidence- centered design	Stage of evidence-centered design
1. Reviewing prior theory and research pertaining to testing issues	No specific components defined	Stage 1—Domain analysis: preliminary synthesis of what is known about what is to be assessed in aviation English

Table 10. (continued)

2. Articulating claims about test takers' language proficiency in all modalities and stating more detailed claims as subclaims	Proficiency paradigm	Stage 2—Domain modeling: incorporation of information from Stage 1 into three components; sketch of potential variables and substantive relationships
3. Listing sources of evidence for each claim	Evidence paradigm	
4. Listing real world tasks for which test takers can provide relevant evidence	Task paradigm	
5. Identifying characteristics that could affect task difficulty	Task paradigm	
6. Identifying criteria for evaluating performance on the tasks	Task paradigm	

In the TDA, the first step involves reviewing prior theory and research connected with testing issues. Over the last two years, the researcher and CSM Ryu, the contact person and domain expert in the 55th ATS Battalion, have met face-to-face and discussed the current research and design of commercial aviation English assessments (e.g., Alderson & Banerjee, 2008; Alderson, 2009; Douglas, 2004; Enright, 2012; Read & Knoch, 2011; Van Moere, Suzuki, Downey, & Cheng, 2011; Werfelman, 2007; Zokic, Boras, & Lazic, 2012). The review of previous research studies helped the researcher with ideas about a potential test format and test environment that would be beneficial for the following five steps.

Articulating claims, the second step of TDA, delineates the proficiency paradigm for the test based on the substantive ability construct that the aviation English test is intended to measure. Such constructs can be expressed as claims one would like to make about test takers (Mislevy et al., 2002, 2003). For example, the construct of the aviation English ability to communicate successfully in ATC tower would be expressed in a claim such as, "The test taker

can communicate effectively by communicating in aviation English to function successfully in the context of an ATC tower in Korean Army Aviation." This general claim can be laid out more specifically through the development of subclaims. Accordingly, a subclaim for aviation English assessment offers a useful means of enunciating a more specific construct. As an example of subclaims for this dissertation study, the test taker can choose appropriate language to accomplish aviation English communication in an ATC tower (such as asking for and receiving a changed flight plan, asking for and giving arrival or departure time, giving information of weather conditions and runway information, etc.).

Listing sources of evidence for claims, the third step of TDA, constitutes the process of defining an evidence paradigm as part of the domain modeling stage. Specifying possible sources of evidence for each claim, the evidence paradigm describes the observations that are necessary to support such claims. In the context of the aviation English assessment, the following aspects of test takers' responses can be identified as the relevant evidence: task accomplishment, appropriate vocabulary, and use of radiotelephony structure.

The remaining steps 4-6 of TDA characterize a task paradigm by identifying real world tasks through which test takers can present relevant evidence, establish task characteristics that could involve task difficulty, and create criteria for performance evaluation. For example, real air traffic control tasks involving English speaking skills consisted of those requiring test takers to comprehend and respond to pilots' request and/or inquiry according to various situations in the ATC tower setting, exchange flight, traffic, or weather information with pilots, and assist pilots in solving problems during the flight. Task characteristics potentially affect difficulty, so task difficulty was identified through task-based needs analysis to help the researcher explore ways to differentiate and sequence assessment tasks according to their difficulty levels as founded in

relevant literature (Norris et al., 1998; Skehan, 1996). Additionally, task-centered rating criteria for evaluating performance on the aviation English tasks were also created based on the findings from the task-based needs analysis survey.

Based on the ECD, the expected outcome is a set of task specifications which include an overall claim and subclaims about what each task is intended to assess. To be more specific, the following task modeling components such as response type, scoring guides, number of questions, task timing, and stimulus information. To provide answers to each step in the TDA, an in-depth task-based needs analysis survey was conducted in the TLU setting; survey findings are provided in the following section.

Task-Based Needs Analysis in TLU Settings

The current section outlines findings from an in-depth task-based needs analysis survey to respond to each step in the Task Design Analysis. The analysis drew upon the first two stages (domain analysis and domain modeling) of Evidence-Centered Design (ECD) process focusing on types of TLU situations and real world tasks in the army aviation English context. Additionally, findings from the needs analysis also provided valuable sources in response to third and fourth stages of ECD process for test development.

The current dissertation study adopted a task-based needs analysis utilizing tasks as the unit of analysis and a task-based performance assessment design (Long, 2005) as opposed to conventional needs analysis frameworks like the target situation analysis (Munby, 1981), the present situation analysis (Richterich & Chancerel, 1980), and the learning centered approach (Hutchinson & Waters, 1987), all of which use linguistic categories (lexical, structural, notional, and/or functional) as the units of analysis. Long and Norris (2000) highlighted two advantages of

a task-based needs analysis over conventional needs analyses: (a) a task-based needs analysis identifies the target language use in real-world situations using the dynamic qualities of the target discourse, while conventional needs analyses provide lists of decontextualized structural items; and (b) the results of a task-based needs analysis can be readily used as authentic input for the task-based language assessment as well as for task-based lessons or course design.

Rationale. As a fundamental step towards the creation and on-going development of a prototype of virtual interactive tasks for aviation English assessment, the current task-based needs analysis study aimed to investigate the characteristics of intended test users and target aviation English tasks and task situations which are the key aspects of the target language use (TLU) situation in an Air Traffic Service (ATS) Battalion in the Republic of Korea. Additionally, this needs analysis study examined task-centered rating criteria by which test takers' performance on the virtual interactive tasks for aviation English assessment can be evaluated.

With reference to the framework of TLU task characteristics (Douglas, 2000) and TBLA (Norris, et al., 1998; Norris, 2001), and Evidence-Centered Design (Mislevy et al., 2002, 2003, 2006), this task-based needs analysis study tried to identify target tasks, task situations in the TLU settings as well as the task-centered rating criteria. The investigation of the target tasks and task situations in the TLU settings focused on perceived importance and analyzed the difficulty of the aviation English tasks. Task-dependent rating criteria, based on what current Korean air traffic controllers believe as excellent, acceptable, and unacceptable (poor) air traffic control performance in the TLU situation, were also investigated to serve as the foundation for the rating rubric in the virtual interactive tasks for aviation English assessment.

Methods. This task-based needs analysis study was utilized to explore the target aviation English tasks and task topics in the TLU situation, including perceived importance and analyzed difficulty of the tasks. It also aimed to examine the rating criteria proposed by experienced air traffic controllers in the TLU situation. To investigate this research question, online open-ended survey responses to participants' beliefs about excellent, acceptable, and unacceptable aviation English task performance were analyzed. The online version of the survey was created and administered using Qualtrics (https://iastate.qualtrics.com/SE/?SID=SV_06aNweVznWqs8KN), and the paper-based version was saved as a document file. Prior to the implementation of the survey, the Korean version of online and paper-based survey was proofread and pre-piloted by CSM Ryu and three Korean graduate students who are currently enrolled in an Applied Linguistics PhD program in U.S.

Participants. The client (contact person) for this task-based needs analysis study was CSM Ji-Wook Ryu (pseudo name), who has been in charge of ATC operational training and testing in the ATS Battalion in the Korean military over the last five years. He has excellent aviation English communication proficiency, though his English pronunciation is slightly influenced by his first language (L1). He has served in the Army aviation as an air traffic controller for about 23 years, and the researcher once worked at the same ATC center with CSM Ryu in Korea from 1997 to 1999 as a military air traffic controller. Access to the research participants was made possible by CSM Ryu. For the online and paper-based survey, a total of 87 Korean air traffic controllers from a military aviation division were recruited for this task-based needs analysis study and 81 successfully completed the survey. The response rate was 93%.

The participants in the task-based needs analysis study can be categorized into two groups: (1) enlisted soldiers (n = 56) and (2) noncommissioned officers (NCO) (n = 25). Enlisted soldiers are all males and were drafted to fulfill the mandatory two-year military service. Their ages range from 19 to 25. Most of them have taken a leave of absence from their undergraduate or graduate studies either in Korea or in other countries. Enlisted air traffic controllers' areas of study at the university are varied and include, but are not limited to, Aviation Operation, Business, Communication, and English. The noncommissioned officers' ages range from 21 to 40 and all of them graduated from the Korea Army Noncommissioned Officer (NCO) Academy after high school or a community college. Fifty-six percent of the noncommissioned officer participants were male (n = 14) and 44% were female (n = 11).

Procedures. One of the initial contacts for this task-based needs analysis study and follow-up dissertation research on the development of the virtual interactive tasks for aviation English assessment was made in the summer of 2012. To discuss the project with the client, CSM Ryu, the researcher visited the headquarters of the ATS Battalion in Korea for two days to obtain approval and support for the project. During the fall of 2012, an IRB (ID: 12-537) application was submitted at Iowa State University and finally approved in December of 2012 as exempt from the requirements of the human subject protections regulations.

This task-based needs analysis was conducted in the spring of 2014. An online and paper-based version of the survey questionnaire was delivered to CSM Ryu in Korea. Eighty-seven recruited military air traffic controllers in Korea responded to either the online or paper-based version of the survey during their on-duty hours. As the survey participants were located in diverse subordinate units and worked under different commands, they were asked to complete the survey on their own under the supervision of CSM Ryu. The data collection took

approximately two weeks, and the scanned paper-based responses were delivered to the researcher by CSM Ryu in March of 2014. Despite the available dual modes of survey data collection, only six participants used the *Qualtrics* online survey, and none of the six actually completed the survey due to their limited access to the Internet for security reasons. Scanned paper-based survey responses were typed in Microsoft Excel and Microsoft Word for analysis.

Data Analysis. To answer the research question about target tasks and task situations in the TLU situation, a task-based needs analysis survey was conducted in Section II of the survey [see Appendix A (English version) and B (Korean version)]. Participants were asked to describe at least five specific aviation English tasks they had performed (or were trained in) for aviation English communication at the ATC tower based on the four language skills (listening, speaking, reading, and writing), and to indicate the level to which they believe it is important to be assessed in the aviation English assessment. Their scanned responses were typed in Microsoft Excel along with the participants' specified importance value. The value of task importance is based on the mean value of participants' response on perceived priority on a six-point Likert scale: 1 – Not a priority; 2 – Low priority; 3 – Somewhat priority; 4 – Moderate Priority; 5 – High priority; 6 – Essential priority.

In addition to the importance index, a task difficulty index was also determined to provide a basis for the predictive utility of aviation English performance assessments. Task difficulty was determined to help the researcher explore ways to differentiate and sequence assessment tasks according to their difficulty levels, as based on a review of literature (Norris et al., 1998; Skehan, 1996). The three variables of central interest in the task difficulty index are (1) code complexity, (2) cognitive complexity, and (3) communicative demand.

The code complexity of a given task addresses the kind of language and information that is involved in successful task performance, and focuses on the (a) range and (b) number of input sources. *Range* indicates the extent to which the code that is inherent in the language of a given task represents a greater or lesser degree of spread. *Number of input sources* represents whether or not the examinee must decode multiple sources of information input. The cognitive complexity of a given task rests on hinges on the amount and kind of information processing that a test taker must engage in to successfully perform the task.

The estimated cognitive complexity of the tasks is based on the two variables: (a) input/output organization and (b) input availability. Input/output organization specifies the extent to which information must be significantly organized for successful task completion. Input availability describes whether or not a test taker is required in some significant way to search for the information upon which a task performance is to be based.

The communicative demand of a given task is established by the type and the number of moderator variables of communicative language activity with the two subsets of (a) mode and (b) response level. *Mode* indicates whether a given task is interpreted to have a productive component for successful task accomplishment. *Response level* denotes the extent to which a test taker must interact with input in a real-time sense.

The difficulty level ranges from zero (least difficult) to six (most difficult) based on the summed numeric value of the six variables (code complexity, cognitive complexity, input/output, input availability, mode, and response level) with zero (generally lower estimated difficulty) or one (generally higher estimated difficulty). To examine the agreement between two raters on the difficulty level coding, interrater reliability was measured as follows.

Table 11

Output of Interrater Reliability for Task Difficulty Coding

	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha	N Agreements	N Disagreements	N Cases	N Decisions
Variable 1 (Rater 1 & 2)	90.91	0.89	0.89	0.89	30	3	33	66

The results of the interrater reliability analysis Cohen's Kappa value is 0.89 with only three disagreements out of 33 cases. This substantial agreement appears to be due to multiple rating calibration sessions between the raters before actual coding.

For the task situations, participants' open-ended responses about the specific aviation English listening, speaking, reading, and writing tasks they had performed at the ATC tower were categorized according to same task situation. Those identified task situation categories were designated as task situations, and involved language skills for each task situation out of the four language skills were also identified. Additionally, the number of involved target tasks and their overall occurrence from the open-ended survey responses were also calculated to present the extent of their significance.

To answer another research question about the rating criteria proposed by experienced air traffic controllers in the TLU situation, participants' open-ended responses about the description of excellent, acceptable, and unacceptable levels of task accomplishment in the TLU situation were elicited through questions 6, 7, and 8 of Section II in the survey questionnaire.

Findings. The next figure and five tables summarize the overall identified target tasks acknowledged by military air traffic controllers, the participants in the TLU situation. In Figure 3, participants' target tasks in the TLU situation, along with the numeric value of perceived importance and difficulty according to the four language skills, are presented. Next, results from

a task-based needs analysis on listening (Table 12), speaking (Table 13), reading (Table 14), and writing (Table 15), and task situation analysis (Table 16) are displayed. Based on responses to open-ended questions, tasks were identified. Eighty-one participants listed at least five tasks they have performed in the TLU situation according to four language skills. A large number of target tasks overlapped and could be integrated into 40 target tasks over the four language skills. Considering the characteristics of ATC communication, it is not surprising that 25 (62.5%) out of 40 target tasks are concerned with listening and speaking skills. Among the four language skills, speaking skills include the most number (n = 14) of diverse ATC tasks, whereas writing skills contain the least number (n = 7) of diverse tasks.

The result shows that task importance value of the target tasks under the category of listening skill is ranked as the highest. This finding is consistent with actual ATC communication in which initial radio transmission is primarily initiated by pilots and the main mission of air traffic controllers is to provide information to pilots in need.

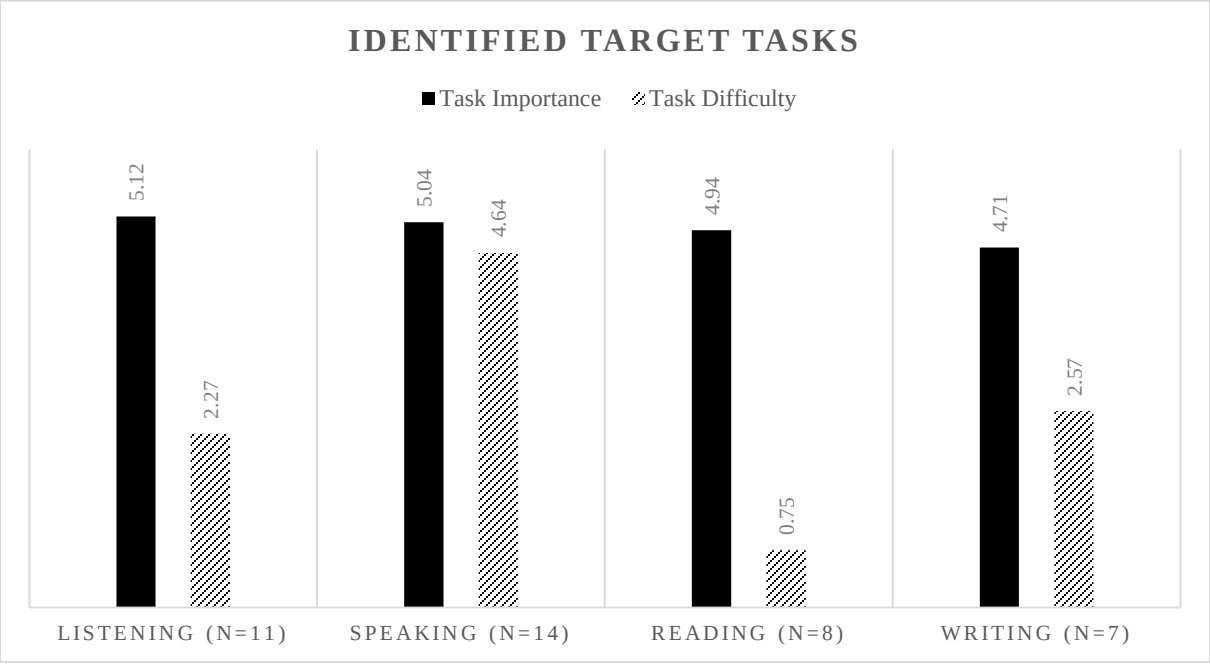

Figure 3. Bar graph showing 81 participants' mean judgments of importance and difficulty of 40 identified target aviation English tasks on a scale from 0 – 6.

Actually, participants judged all four skills to be almost equally important. The only real variation was in difficulty. The Interestingly, compared to high task importance level in listening, the analysis result of 11 for listening target tasks shows that the difficulty level of listening tasks is considerably low. Comparing all of the four skills, the receptive language skills, listening and reading, are regarded as less difficult than the productive language skills, speaking and writing. Overall, target speaking tasks were perceived as significantly important and difficult, which can provide valuable input for task design and development.

Listening Tasks. Though task importance and task difficulty were intended to assess different characteristics of target tasks, the two index values seem to be consistent to some extent. Among the 11 listening tasks in Table 12, *Listen to PIREP* (pilot report) was perceived and reported to be the most critical target task, as this task was mentioned 11 times. In contrast with other reports and flight controls, PIREP is not planned and transmitted unexpectedly. PIREP includes hazardous weather conditions and forest fires observed by pilots in flight, so failure in understanding PIREP could lead to a tremendous loss. Interestingly, all of the listed listening tasks are not independent tasks, but should be followed by air traffic controllers' utterances providing requested information or approval.

Table 12

Listening Target Tasks (N=11)

	Imp. index	Diff. index	Code		Cognitive complexity		Communicative demand	
			range	input sources	in/out organiz.	input avail.	mode	resp. level
Listen to PIREP (weather, forest fire) (11)	5.5	3	+	+	-	-	-	+
Listen to pilots' request of weather information (4)	5.5	3	+	+	-	-	-	+
Listen to pilots' current position report (2)	5.5	3	+	+	-	-	-	+
Listen to pilots' landing/departure request (9)	5.2	2	-	+	-	-	-	+
Listen to changed flight plan (9)	5.2	3	+	+	-	-	-	+
Listen to emergency (procedure) landing (5)	5.2	3	+	+	-	-	-	+

Table 12. (continued)

Listen to pilots' call for inbound to & outbound from the control zone (9)	5.1	2	-	+	-	-	-	+
Listen to pilots' transition request (6)	5	1	-	-	-	-	-	+
Listen to emergency evacuation (3)	5	3	+	+	-	-	-	+
Listen to pilots' taxi request (3)	4.7	1	-	-	-	-	-	+
Listen to American pilots' plain English (5)	4.4	1	-	-	-	-	-	+

Notes. The numbers in parentheses refer to the number of mentions made in the open-ended questions. Imp. = Importance; Diff. = Difficulty; Resp. = Response.

Speaking Tasks. As indicated in Figure 3, as many as 14 target speaking tasks were identified (see Table 13). In agreement with the most important listening task, *Listen to PIREP*, participants noted *Provide PIREP to other pilots* as most important and also analyzed this as the most difficult target task. This finding can be adopted in the target task scenario development so that the prototype VIAET can be more authentic and interactive based on the expected test takers' needs. Diverse speaking tasks can be categorized under three major task topics--departure, landing, and transition -- which are agreed upon in the TLU situation. Another intriguing finding from the survey results is that identical task type (Long, 1985; Long & Norris, 2000) can be executed in various situations. For example, the most frequently mentioned target task, *Provide weather information to pilots*, is actually performed in diverse contexts, including departure, landing, and transitioning. This may suggest that mastery of a given task type is critical, but appropriate application and modification of the task type into a variety of task situations should be more demanding and relevant for advanced air traffic controllers.

Table 13

Speaking Target Tasks (N=14)

	Imp. index	Diff. index	Code		Cognitive complexity		Communicative demand	
			range	input sources	in/out organiz.	input avail.	mode	resp. level
Provide PIREP to other pilots (4)	6	6	+	+	+	+	+	+
Provide relayed information from ground controls to pilots (4)	6	5	+	+	+	-	+	+
Provide landing instruction to pilots (13)	5.5	5	-	+	+	+	+	+
Provide holding information when traffic is too crowded (11)	5.5	5	-	+	+	+	+	+
Provide traffic information to transitioning pilots (13)	5.3	5	-	+	+	+	+	+
Provide departure instruction to pilots (8)	5	5	-	+	+	+	+	+
Provide emergency procedure (6)	5	5	+	+	-	+	+	+
Provide warning to unauthorized aircrafts (2)	5	5	+	+	-	+	+	+
Provide weather information to pilots (15)	4.9	5	+	+	-	+	+	+
Provide traffic information to pilots in local flight (6)	4.8	5	-	+	+	+	+	+
Provide taxi / ramp instruction (5)	4.8	5	-	+	+	+	+	+
Explain reasons for aircraft holding (2)	4.5	4	+	-	+	-	+	+
Request PIREP from pilots (2)	4.5	2	-	-	-	-	+	+
Provide NOTAM information to pilots (5)	3.8	3	-	-	+	-	+	+

Notes. The numbers in parentheses refer to the number of mentions made in the open-ended questions. Imp. = Importance; Diff. = Difficulty; Resp. = Response.

Reading Tasks. Though the task difficulty index is the lowest among the four language skills, participants specified as many as eight target tasks concerning reading skills (see Table 13). Among the tasks, *Read and understand NOTAM* (a notice to airman) is ranked as the most important task in reading skills. As the information listed in NOTAM is especially critical in air safety, air traffic controllers are expected to decode all of the symbols, and abbreviated codes. Regarding the number of mentions made in the survey responses, *read and understand weather information symbols & abbreviation symbols* were highlighted by the participants the most frequently. This result represents what specific target resources, such as NOTAM, ATC phraseology, flight plan, weather symbols, and FLIP (Flight Information Program), are essential for participants' ATC task performances in the real world. Therefore, those resources can be adopted as authentic target task instruments.

Table 14

Reading Target Tasks (N=8)

	Imp. index	Diff. index	Code		Cognitive complexity		Communicative demand	
			range	input sources	in/out organiz.	input avail.	mode	resp. level
Read and understand NOTAM (8)	5.9	2	+	+	-	-	-	-
Read standard (abbreviated) ATC phraseology (6)	5.3	1	+	-	-	-	-	-
Read and understand flight plan (13)	5.2	2	+	-	-	-	-	+
Read and understand abbreviations for aircraft parts and maintenance (2)	5	1	+	-	-	-	-	-
Read and understand weather information symbols & abbreviation (24)	4.6	0	-	-	-	-	-	-
Read and understand FLIP (11)	4.6	0	-	-	-	-	-	-
Read and understand AIRAD (2)	4.5	0	-	-	-	-	-	-
Read and understand airbase information (5)	4.4	0	-	-	-	-	-	-

Notes. The numbers in parentheses refer to the number of mentions made in the open-ended questions. Imp. = Importance; Diff. = Difficulty; Resp. = Response.

Writing Tasks. Contrary to participants' general belief about ATC tasks, numerous writing target tasks were revealed from the survey (see Table 15). Similar to the findings from reading target tasks, the task instruments in writing tasks include PIREP, weather information, flight plan, and NOTAM. However, what is worth noting is the relatively higher level of difficulty index. The most difficult tasks in writing skills, *Take notes of changed flight plan during a flight*, as well as other difficult tasks, *Take notes of PIREP* and *Take notes of ground controls request (information) for pilots,* are not merely writing tasks, but also tasks which are synchronously performed with listening. If test takers fail to listen to the target transmissions, they cannot take notes correctly. Given the importance and difficulty index of the writing tasks, such note taking tasks need to be included in the VIAET to challenge the test takers and differentiate their task accomplishment levels.

Table 15

Writing Target Tasks (N=7)

	Imp. index	Diff. index	Code		Cognitive complexity		Communicative demand	
			range	input sources	in/out organiz.	input avail.	mode	resp. level
Take notes of PIREP (4)	5.3	4	+	+	-	-	+	+
Take notes of weather information (9)	5	3	+	-	+	-	+	-
Write daily ATC log (8)	4.8	1	-	-	-	-	+	-
Write a strip (flight plan) (9)	4.7	1	-	-	-	-	+	-
Take notes of changed flight plan during a flight (7)	4.7	5	+	-	+	+	+	+
Take notes of NOTAM (3)	4.3	1	-	-	-	-	+	-
Take notes of ground controls' request (information) for pilots (5)	4.2	3	+	-	-	-	+	+

Notes. The numbers in parentheses refer to the number of mentions made in the open-ended questions. Imp. = Importance; Diff. = Difficulty; Resp. = Response.

Task Situations. The analysis of target tasks according to four language skills reveals the characteristics of individual target tasks and how target tasks from one English skill are interrelated with other skills to complete target tasks in various task situations. Considering the interrelated feature of aviation English tasks in TLU settings, the analysis result of task situations synthesizing four language skills is presented in Table 16. Based on the analysis of 275 target tasks in the open-ended questions across the four language skills, 14 task situations were identified. Findings from the analysis show what specific language skills were involved with the specific task situations. The total occurrence presents the portion of frequency of individual task situation out of the entire 275 target tasks from the open-ended survey questions. Approximately 19% (n = 52) of target tasks were about *weather* followed by *flight plan*, *transition, landing, PIREP, traffic flow,* and *departure*. The 14 task situations can be further categorized into three primary task situations of landing, departure, and transition, and other task situations can play roles in supporting those three primary task situations. Task situations and sequences were identified by the findings from the task-based needs analysis survey, ATC experts' consensus,

and ATC manual (AAS, 2012) analysis. Specific target tasks corresponding to each task situation were further incorporated in an authentic way based on the ATC training manual, experts' consensus, and the researcher's ATC experience (see Appendix D). This result clearly highlights how the situational portion of ATC communication is shaped and provides authentic resources for task design and development.

Table 16

Summary of Task Topics across Language Skills

Task Topic	Language Skills	Total Task Occurrence
Weather	L,S,R,W	52 (19 %)
Flight Plan	L,S,R.W	49 (18 %)
Transition	L,S	28 (10 %)
Landing	L,S	22 (8 %)
PIREP	L,S,R,W	21 (8 %)
Traffic Flow	L,S	19 (7 %)
Departure	L,S	16 (6 %)
Emergency Flight	L,S,R,W	16 (6 %)
NOTAM	S,R	15 (5 %)
Layover Information	L,S.R,W	14 (5 %)
Aviation Phraseology	L,S,R,W	8 (3 %)
Flight Log	R,W	8 (3 %)
Colloquial English	L,S	5 (2 %)
Position Report	L,S	2 (1 %)

Notes. L = Listening; S = Speaking; R = Reading; W = Writing; the numbers in parentheses refer to the percentage out of the entire 275 target tasks.

Alternative Rating Rubric. Given the multiple test use – decision making and diagnostic – expected in the target context, it is highly important to design alternative rating rubrics, to be used in addition to the ICAO's LPRs rating scale, which could provide useful test results for multiple inferential purposes. This section describes a potential alternative rating rubric proposed

by experienced air traffic controllers, focusing on their opinions about excellent, acceptable, and unacceptable air traffic control performance in the TLU situation. Participants' open-ended survey responses were analyzed as task-centered rating criteria; to identify the rating criteria, participants' descriptions of excellent, acceptable, and unacceptable level of task accomplishment were elicited in questions 6-8 of the task-based needs analysis survey questionnaire.

The following are two representative air traffic controller participants' open-ended responses about excellent, acceptable, and unacceptable task performance in the context of ATC tower:

Participant A (a private first class controller)

Excellent: provide pilots with requested information quickly with accurate pronunciation; fully aware of local flight procedures; deal with emergency quickly and accurately

Acceptable: provide pilots with requested information during routine air traffic control; however, lack of ability during emergency

Unacceptable: be slow in responding to pilots during routine air traffic control; fail to deal with emergency air traffic control

Participant B (a sergeant first class controller)

Excellent: understand pilots' request accurately; respond to pilots' request using standard ATC phraseology; handle heavy traffic smoothly; be able to predict upcoming situations;

Acceptable: deal with basic local flight control without many mistakes; comprehend most of the pilots' requests

Unacceptable: fail to understand most of the pilots' requests; provide pilots with wrong information; fail to understand the traffic flow of local flight

The participants described evaluative criteria by focusing not only on various linguistic elements, but also on various performance elements to differentiate the extent of task accomplishment: excellent, acceptable, and unacceptable. Over diverse task situations, participants' descriptions about the extent of task accomplishment were fairly consistent. Table 16 shows an example of task-centered rating criteria for excellent-, acceptable-, and unacceptable-level task accomplishment in the landing situation. Task-centered rating criteria in each level correspond to one another.

Table 17

An Example of Task-Centered Rating Criteria for Landing-Related Tasks

Rating Level	Task Situation - Landing
Excellent	"able to understand all the information requested by pilots in landing procedure"
	"able to identify air traffic flow around the air space clearly"
	"able to accurately command standard aviation English phraseology for landing"
	"able to transmit traffic and weather information to pilots quickly and accurately"
	"able to approve pilots' request during landing correctly and fast"
Acceptable	"able to understand most information requested by pilots in landing procedure"
	"able to identify most part of air traffic flow around the air space"
	"able to command most of the standard aviation English phraseology for landing"
	"able to transmit most traffic and weather information to pilots"
	"able to approve pilots' request during landing correctly, but delayed"
Unacceptable	"fail to understand information requested by pilots in landing procedure"
	"fail to identify air traffic flow around the air space"
	"fail to command standard aviation English phraseology for landing"
	"fail to transmit traffic and weather information to pilots"
	"fail to approve pilots' request during landing"

Actually, all of the excellent-level criteria from the participants' open-ended responses entail more extensive and more specific criteria than those found at the other performance accomplishment levels. In other words, task-centered rating criteria at the excellent level could be considered any target aviation English task performance that is particularly accurate, clear, or quick, whereas an acceptable-level task performance is considered any performance that obviously surpasses the minimum components that would be necessary for successful task accomplishment without exceedingly positive adjectives or with rather negative adjectives. Contrary to excellent and acceptable level descriptions, unacceptable-level task performance indicates a target task performance that does not result in task accomplishment, with the use of "*fail to*" in the description.

Initially, it was expected that participants' responses about task-centered rating criteria would be specific enough to differentiate the level of task performance in each target task. Interestingly, however, participants' responses about task-centered rating criteria concentrated on the level of task situation rather than sub-component target tasks that are critical to the successful accomplishment of the task situations (e.g., ATC communication in aircraft landing). Accordingly, participants' task-centered rating criteria were naturally clustered together under the same task topic situations. For example, in ATC communication involving a pilot who is trying to land, numerous tasks are encompassed in the successful accomplishment of the landing procedure. As reported in Table 13, for successful landing, on-going listening and speaking tasks are required to be performed until there is a full stop on the ramp.

The finding is quite interesting and useful in the process of task-centered rating criteria development. It is important to consider and assess individual task completion in order to make inferences about test takers' successful task accomplishment in the TLU situation. However,

reflecting the findings from the current task-based needs analysis study, the scope of the task-centered rating criteria should entail not only successful accomplishment of task topics, but also the difficulty level of individual tasks that constitute those task situations. Especially, for logistical reasons and for the sake of test users, test takers, and test raters in the military, the design of task-centered rating rubrics was based on the identified target task situations with key target tasks under the scope of each task situation.

Outcomes from the Task Design Analysis. The outcome of TDA brings the researcher closer to defining the aviation English tasks by synthesizing necessary data from the task-based needs analysis results. Table 18 exemplifies the aviation English test task used to construct the task shells.

Table 18

Outcome of Task Design Analysis for Aviation English Assessment

Step in task design analysis	Outcome for the aviation English Assessment
Reviewing previous research and other relevant assessments	Ideas about aviation English proficiency and potential test tasks
Articulating claims and TLU situations	Claims: The test taker is able to communicate in aviation English which is needed to function properly in the context of an air traffic control tower. TLU situations: The test taker can generate / comprehend aviation English effectively to / by native and non-native English speaking pilots. The test taker can select appropriate plain language and aviation phraseology to carry out aviation English communication (such as giving directions; receiving requests; asking for and giving information; asking for clarification; etc.).
Listing sources of evidence	Appropriate completion of task, relevant aviation phraseology and use of correct aviation English structure
Listing real world tasks in which test takers can provide relevant evidence	Listen to pilots' request of weather information; provide PIREP (pilot report) to other pilots; read standard (abbreviated) ATC phraseology; write a strip (flight plan); etc. (see Table 11, 12, 13, 14)

Table 18. (continued)

Identifying aspects that would affect task difficulty	To identify performance difficulty characteristics, six variables – range, input source, in/out organization, input availability, mode, and response level – were measured for each target task. (see Table 11, 12, 13, 14)
Identifying criteria for evaluating performance on the tasks	Language-centered rating criteria - ICAO's language proficiency requirements focusing on comprehension, fluency, interactions, pronunciation, structure, and vocabulary (see Appendix C); task-centered rating criteria – focusing on the level of accomplishment of given tasks (see Appendix D)

The information, which is presented for aviation English assessment in Table 18, is actually used to construct a task shell, a template for generating parallel test tasks and is made up of two primary components: a brief summary of what the target task is supposed to measure and a task model.

An example of a task shell for aviation English assessment is shown in Table 19. In the first column, a claim is provided according to the claims identified for listening comprehension during Step 2 of the TDA shown in Table 10. In this example, the claim "Test takers can comprehend radio transmission of a flight plan produced by native and proficient non-native English speaking pilots" reflects the TLU situation 1 in Table 19. The phrase above in the column title of Claim in the What is being measured, "Aviation English ability comprehending a flight plan transmitted by a pilot over a radio" summarizes what the tasks developed from this this shell are intended to measure. Under the claim, the aspects of the test takers' responses or comprehension that are being assessed are listed. The task model is also prepared with the specifications for the fixed elements, variable elements, rubric and list of variants. These task shells are then used, in turn, to generate test tasks.

Table 19

An Example of Task Shell for a Flight Plan Listening Task

What is being measured	Task model			
	Fixed elements	Variable elements	Rubric	Variants
Aviation English ability comprehending a flight plan transmitted by a pilot over a radio Claim: Test takers can comprehend radio transmission of a flight plan produced by native and proficient non-native English speaking pilots. Measurement: Analytic evaluation of: Comprehension, structure, interaction Holistic evaluation of: Dictation of the flight plan	1. Nature of the task – Demonstrate ability to comprehend pilots' radio communication Features of the radio communication to be heard: - short radio transmission (3-10 sec) - context combined 2. Order of item elements - Task-specific contexts will be given to the test taker prior to a pilot's radio transmission. As soon as the test taker hears the pilot's call or request, the test taker will respond immediately.	1. Type of task situations 2. Purpose of the radio transmission	See Appendix C, D	Including: - Request of weather information - Request of taxi instruction - Request of runway information - Request of transition - Report of emergency - Report of a changed flight plan - Report of leaving control zone

Overall, Mislevy's four-stage ECD process for the virtual interactive tasks for aviation English assessment presented in Table 8 was referred to in developing the example of a task shell shown in Table 19. Once such a task shell and some sample tasks were created, the shell was evaluated by CSM Ryu, the TLU domain expert. On-going exploration of the task situations and task difficulty integrating all of the four aviation English listening, speaking, reading, and writing skills resulted in the end draft of final blueprint (see Table 20) and specification for the aviation English ability measures.

Table 20

Final Blueprint for the Virtual Interactive Tasks for Aviation English Assessment

Final Blueprint – Virtual Interactive Tasks for Aviation English Assessment	
Stimulus	Items per task situation
1. Colloquial communication	4 questions - service period; perception about ATC job; favorite Korean food; places to eat out
2. Changed flight plan	3 expected responses including mixed language skills – listening, writing, and speaking
3. Terminal information	1 expected response including mixed language skills – listening, speaking, and reading
4. Departure procedure	6 expected responses including mixed language skills – listening, speaking, and reading
5. Transition procedure	4 expected responses including mixed language skills – listening and speaking
6. Arrival procedure	5 expected responses including mixed language skills – listening and speaking
7. Arrival and departure procedure	10 expected responses including listening, speaking, and reading
	Total test time – 20 minutes

Initial analysis from the TDA and task-based needs analysis outcomes guided the researcher to focus on the three independent task situations: departure, landing, and transition. The task situations in the first column are based on the findings regarding task topics across language skills (see Table 16). Yet, after discussions and piloting with TLU experts, including CSM Ryu, the researcher paid more attention to revision of the task situations and transition from one situation to another situation to enhance authenticity. Otherwise, test takers may perceive each independent task situation as less authentic and less cognitively situated in the simulation settings, which could result in impairing validity of the test in the end. Though the first task situation, colloquial communication, was ranked as one of the least frequently occurring task situations, due to increasing combined military ATC exercises between the U.S. and Korean air traffic controllers, CSM Ryu strongly suggested that a task about colloquial communication, which is different from plain English or language for aviation communication,

should be included in the test to meet the test users' needs. To enhance the authenticity, the entire test's situations were sequenced in a realistic and natural order with the expectation that test takers would not feel awkward during the task performance and be more emerged into the virtual test task settings.

The final outcome of the TDA for virtual interactive tasks for aviation English assessment is summarized in Table 21, which functions a set of task specifications for the seven task situations. These specifications (see Table 21) include an overall claim and TLU situations about what each task situation is intended to assess.

Table 21

Summary of Specifications for Virtual Interactive Tasks for Aviation English Assessment

Aviation English Claim	Test takers can communicate in Aviation English to function appropriately in the context of ATC tower.						
TLU Situations	Test taker can communicate in colloquial English	Test taker can comprehend flight plan	Test taker can generate Aviation English for terminal information	Test taker can direct departure procedure	Test taker can respond to transition request	Test taker can direct landing procedure	Test taker can generate effective English to process synchronous requests of landing, departure
Nature of Aviation English task	Respond to short questions based personal experience	Comprehend changed information	Respond to requests based on written information	Respond to request based on training and experience	Respond to short request based on experience	Respond to request based on training and experience	Propose a solution based on a problematic situation
Scoring rubric	Analytic 0-6 Task-centered 0-2:	Analytic 0-6 Task-centered 0-2:	Analytic 0-6 Task-centered 0-2:	Analytic 0-6 Task-centered 0-2:	Analytic 0-6 Task-centered 0-2:	Analytic 0-6 Task-centered 0-2:	Analytic 0-6 Task-centered 0-2:
Number of questions	4	3	1	6	4	5	10
Nature of stimulus material	Simulated airbase with listening stimulus – voice from an avatar	Simulated ATC tower with the view of a grounded helicopter; voice from the pilot	Listening stimuli: a pilot's voice that requests terminal information	Listening stimuli: 6 short requests for departure; reading passage about terminal information;	Listening stimuli: 4 short requests for transition	Listening stimuli:5 short request for landing instruction; interactive helicopters in the air	Listening stimuli: 10 short request for emergency landing, NOTAM, and departure;

Table 21. (continued)

							interactive helicopters on the ground, in the air
				interactive helicopter			
Prep time	3-5 sec.	3-5 sec.	3-5 sec.	3-5 sec.	3-5 sec.	3-5 sec.	3-5 sec.
Response time	2 min.	2 min.	1 min.	2 min.	2 min.	3 min.	4 min.
Total time	Approximately 16 minutes for 7 task situations						

As demonstrated earlier, seven TLU situations were identified which can provide support for the overall claim that a test taker can communicate in Aviation English to function appropriately in the context of an ATC tower. These TLU situations were sequenced in such order so as to reflect authentic aviation English tasks and feedback from the domain experts. Additionally, the complexity and difficulty of the target task situation was also considered by placing less challenging tasks in the beginning and the most challenging task at the end so the tests gradually becomes more complex and test takers could be more comfortable in the simulation environment with less stress or anxiety during task completion. A complete set of task situations with specific scripts between pilots and the test taker is provided in Appendix D.

Simulating Target Test Tasks into a Virtual World (Second Life)

This section describes the process of simulating target test tasks and task situations in a virtual world. The chronological order of the actual design and development of test tasks in Second Life is as shown in Table 22.

Table 22

Summary of the Process of Creating Test Tasks in Second Life

Process	Time Period	Description
Conducting TLU analysis	May 2012 ~ Aug. 2013	Target language use analysis with documents of ATC training manuals, reference books, observation of ATC performance in the military base.
Conducting task-based needs analysis	Mar. 2014~Apr. 2014	Task-based needs analysis was conducted and preliminary target tasks were identified.
Recording audio files	Jul. 2014 ~ Aug. 2014	Audio recording by actual air traffic controllers in the TLU setting was completed (one non-native and one native English speaking controllers were recruited).
Simulating in Second Life	Jul. 2014 ~ Dec. 2014	1. While audio recording was conducted, Dr. Sadler and the researcher started to simulate the target military airbase into Second Life comparing satellite images and photographs about the target settings. 2. After the simulation of ATC tower and airbase was completed, audio files were embedded into the simulation environment.
Piloting	Jan. 2015	Pilot test was conducted in the military with actual ATC controllers to identify issues of the tasks and simulation environments.
Revising simulation	Jan. 2015	Revision on the simulation environment and tasks was conducted with Dr. Sadler.

The initial phase of the task simulation started with the analysis of the target domain where the researcher worked as an air traffic controller from 1998 to 1999. Though the researcher had previously experienced TLU situations, it was crucial to receive updated

information and be aware of the most current issues in aviation English training and testing from current TLU experts. As the researcher has been in touch with CSM Ryu since retirement from the ATS Battalion in December 1999, the researcher had been informed about the issues and needs for innovation of aviation English assessment. To ensure compliance and support from the ATS Battalion, the researcher visited the airbase in May 2012 and was able to observe the current practice of ATC and obtain ATC resources, such as a training manual (AAS, 2012) and reference books, for in-depth document analysis on the TLU situation.

As demonstrated in previous sections, a task-based needs analysis study was conducted to identify test tasks for this dissertation study. As soon as the target task situations and test tasks were ready, the task scripts were produced based on the standard procedures as described in the military ATC manual, findings from the task-based needs analysis about ATC procedures, and experts' consensus. Then the task scripts were sent to the ATS Battalion for audio recording at a real ATC tower, including authentic helicopter rotor blade noise during the recording. The voice recording was done by a male non-native English speaking ATC sergeant at the ATS Battalion and was also done by a male U.S. ATC sergeant stationed in Korea.

During the recording of the audio files, the researcher invited Dr. Randal Sadler, one of the most prominent researchers on virtual environments, especially Second Life, to be a member of the researcher's dissertation committee. Dr. Sadler emphasized authenticity of the simulation environments, including such details as the design of military uniform, color of the face, items displaced in the tower, and so on.

Figure 4 presents two images displaying the front view in the ATC tower; the left picture shows the front view in the actual ATC tower in the target context and the right image shows a screen capture image of the ATC tower in the SL testing environment.

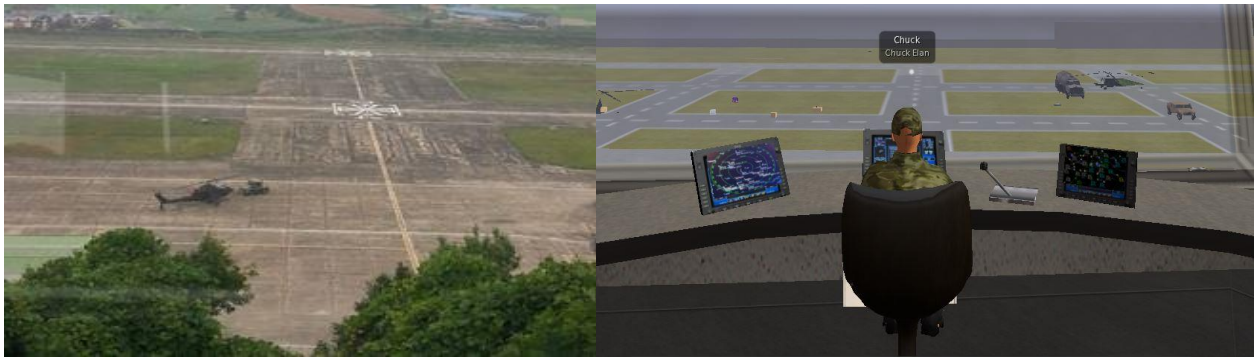

Figure 4. Actual and emulated images of the front view of the ATC tower

Once the simulated airbase, including the ATC tower, helicopters, and user avatars, were designed and developed, it was fairly manageable to design a variety of authentic target tasks controlling the mission of a flight, the number of aircrafts, and the complexity of ATC tasks. Such manipulation and implementation of authentic target tasks in SL held the capability of saving a considerable amount of time, as tasks could be easily duplicated and edited. After visual simulation was completed, MP3 files of the target task scripts were embedded into the Second Life system so for convenient and fast test implementation.

In January 2015, a pilot test in the simulated ATC tower in Second Life was conducted with two Sergeant First Class ATC controllers in the ATS Battalion in Korea. Over two-day pilot testing period followed by discussion with TLU experts on the base, several issues regarding the scripts and simulation settings were identified. A number of missing transmissions in the script were added and additional audio recording was conducted right after the piloting. The issue of synchronizing helicopter movement and the development of remote control buttons of audio files in the simulated ATC tower were also improved with the technical support of Dr. Sadler.

Utilizing a systematic ECD framework allowed the provision of more solid backing for the part of the validity argument pertaining to test design and development. What is required to establish a comprehensive validity argument requires documenting test design and development

as an important phase of the validity process. The role of a test developer, especially in this dissertation study, is not limited to the justification of the appropriateness of test design, but also providing relevant evidence for appropriate test use (Shepard, 1997). The following section discusses an interpretive argument for the virtual interactive tasks for aviation English assessment including the approach to the interpretive and validity argument.

An Interpretive Argument for the Virtual Interactive Tasks for Aviation English Assessment

An interpretive argument indicates the proposed interpretations and uses of test results and serves as the first step in developing a validity argument (Kane, 1992, 2001, 2006). The interpretive argument describes the reasoning inherent in the proposed test score interpretations and uses based on a sequenced framework extending from the grounds to the claims by inference. It is intended as a solid starting point for validation of test score interpretations and uses. This section presents (a) an overview of test intended interpretations, uses, and consequences, (b) the approach to validation, and (c) the interpretive argument.

An Overview of Test Interpretations, Uses, and Consequences

The prototype of virtual interactive tasks for aviation English assessment has been developed as a measure of Korean military air traffic controllers' aviation English ability. This prototype aviation English test development and validation study is expected to contribute to test users' decision about test takers' assignments by spread out the novice and experienced air traffic controllers with respect to their demonstrated level of aviation English communication abilities and to provide diagnostic feedback to the air traffic controllers to improve their aviation English

communication skills. Test takers' task performance in the virtual interactive tasks for aviation English assessment is expected to provide predictive evidence to test users with regards to their decisions about the test takers. The consequences of using the virtual interactive tasks for aviation English assessment and of the decisions that are made based on them will serve to enhance U.S. and Korean army aviation forces' capacity in Korea. This project is certainly timely considering the recent agreement between the stationed U.S. and Korean militaries to share both army ATC frequencies, an agreement that demands more accurate and fluent aviation English communication skills and ability from Korean air traffic controllers. Test score interpretation and use is expected to be beneficial for all test users, including CSM Ryu and other ATS Battalion officers in charge of ATC training and personnel administration and other stakeholders, Korean army air traffic controllers, and U.S. and Korean army pilots. The underlying benefit of the virtual interactive tasks for aviation English assessment validation is to raise awareness of the administrators and instructors in the Army Aviation School where actual ATC training is conducted so that they can adopt and apply task-based ATC training and assessment in a virtual environment based on the analysis of authentic ATC communication between air traffic controllers and pilots. Furthermore, the evaluative feedback from the implementation and validation of the prototype virtual interactive tasks for aviation English assessment will be a valuable resource for test developers to expand, modify, and improve the prototype virtual interactive tasks for aviation English assessment as a more reliable and valid assessment tool than conventional paper-based aviation English tests in the target contexts.

Approach to Validation

According to Kane (2006), validating a proposed interpretation or use of test scores is

evaluating the rationale for this interpretation or use, so the evidence necessary for validation depends on the interpretations and uses that are claimed. Thus, validation requires a statement of the proposed interpretations and uses by employing two kinds of argument – an interpretive argument and a validity argument. An interpretive argument references the proposed interpretations and uses of test results by systematizing the network of inferences and assumptions from the observed performance to the conclusions and decisions based on the test takers' performances (Kane, 2006). Throughout the test development and follow-up validation process, Kane's (2006) and Chapelle et al.'s (2008) approaches have been referred to as guiding models for the interpretive argument aspect of this dissertation study. As Kane's interpretive argument enables the reasoning inherent in the proposed interpretations and uses, his sequenced framework extending from the grounds to the claims with a chain of inferences can be a solid starting point for both test developers and test evaluators.

Chapelle et al. (2008) highlights target domain analysis, which is a crucial phase for the interpretive argument study, as this dissertation study considers the importance of aviation English in English for Specific Purposes (ESP). Furthermore, Chapelle et al.'s (2008) approach is more ideal after adding the concept of test use inference, an element missing in some of Kane's (2006) examples. Chapelle et al.'s (2008) inclusion of the test use sequence has the potential to help test evaluators like the researcher better adopt and apply frameworks for test development purposes. Toulmin's (2003) approach to practical reasoning with a claim, data, warrant, counterclaim, rebuttal, and rebuttal backing is the basis for an interpretive argument. The structure of Toulmin's practical argument that the researcher applied in this study can be perceived as illustrated in Figure 5.

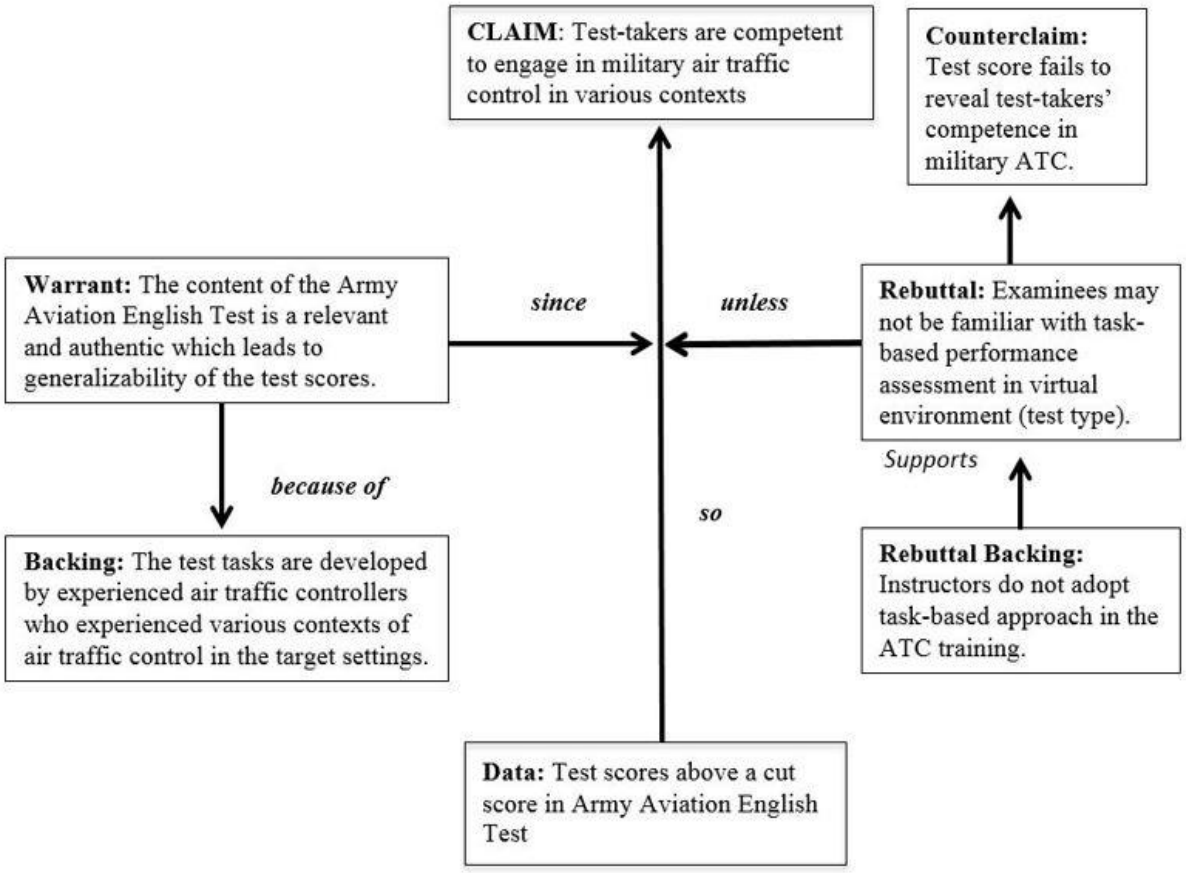

Figure 5. Structure of practical argument for Korean army aviation English tests

In this structure of an interpretive argument about virtual interactive tasks for aviation English assessment, the claim indicates the conclusions drawn about test takers based on data or observations about test takers' performance (data). The data is made up of empirical observations (e.g., test scores) on which the argument is built. A warrant links the data to a claim, legitimizing the claim by showing the data to be relevant. Backing for an argument gives additional support to the warrant. A rebuttal suggests a counter-argument to the claim.

Based on the test score data to be collected from virtual interactive tasks for aviation English assessment in the Korean Army ATS Battalion, a claim can be made that test takers who score above acceptable level of task performance on the virtual interactive tasks for aviation English assessment test are competent to engage in various army air traffic control contexts. The

warrant that the researcher can use to justify the claim is that the content of the virtual interactive tasks for aviation English assessment are relevant and authentic, which leads to generalizability of the test performance. This warrant can be supported by the evidence which suggests the authenticity and relevance of the test items. The backing can be drawn from the fact that the aviation English test tasks are developed by experienced army pilots and air traffic controllers who have experience in various contexts of air traffic control.

However, there exists a condition under which the warrant may not apply. A possible rebuttal in this context could be test takers' unfamiliarity with task-based performance assessment in a virtual world (Second Life). In this hypothetical instance, the rebuttal could be rejected if it has no backing, which may suggest counter evidence in that instructors in the Army Aviation School in Korea adopt and apply task-based approaches in the ATC training or test takers have an adequate period of adjustment training in Second Life before they take the assessment.

The interpretive argument expanded from Toulmin's practical argument includes several inferences that require explanation. As introduced in Chapelle et al. (2008), domain description links test takers' performances in the target domain to the observations of performance in the test domain. Evaluation links the observations of performance on test tasks with observed scores reflecting test takers' target language abilities. Generalization links the scores assigned to the test takers' performance with their expected score on similar test tasks under similar conditions.

Though the virtual interactive tasks for aviation English assessment are a prototype test and not a complete set of test battery, this dissertation research, based on the interpretive argument framework, aims to include the entire consideration of inferences - domain description, evaluation, generalization, explanation, extrapolation, and utilization.

The Interpretive Argument

This section outlines the interpretive argument which consists of the claims, warrants, and assumptions. The major sources of the interpretive argument framework were Chapelle et al.'s (2008) frameworks of inferential links, while Bachman and Palmer's Assessment Use Argument (AUA) flowchart (Bachman and Palmer, 2010, p. 91) was also referred to in developing a more solid, overarching framework including the target domain as well as the intended / actual decisions and consequences (see Figure 6).

In Figure 6, a specific air traffic controller's expected virtual interactive tasks for aviation English assessment performance is used to demonstrate the grounds and claims of the interpretive argument as well as the inferences which link the grounds and the claims. The consequences as the conclusion for this argument are supported with backing for the warrants in the chain of inferences leading to the conclusion. The interpretive argument includes seven inferences (Domain description, Evaluation, Generalization, Explanation, Extrapolation, Utilization, and Implication) and eight grounds/claims. This interpretive argument consisting of different types of inferences provides guidance as to the types of research needed for this dissertation study, which will be presented after this section.

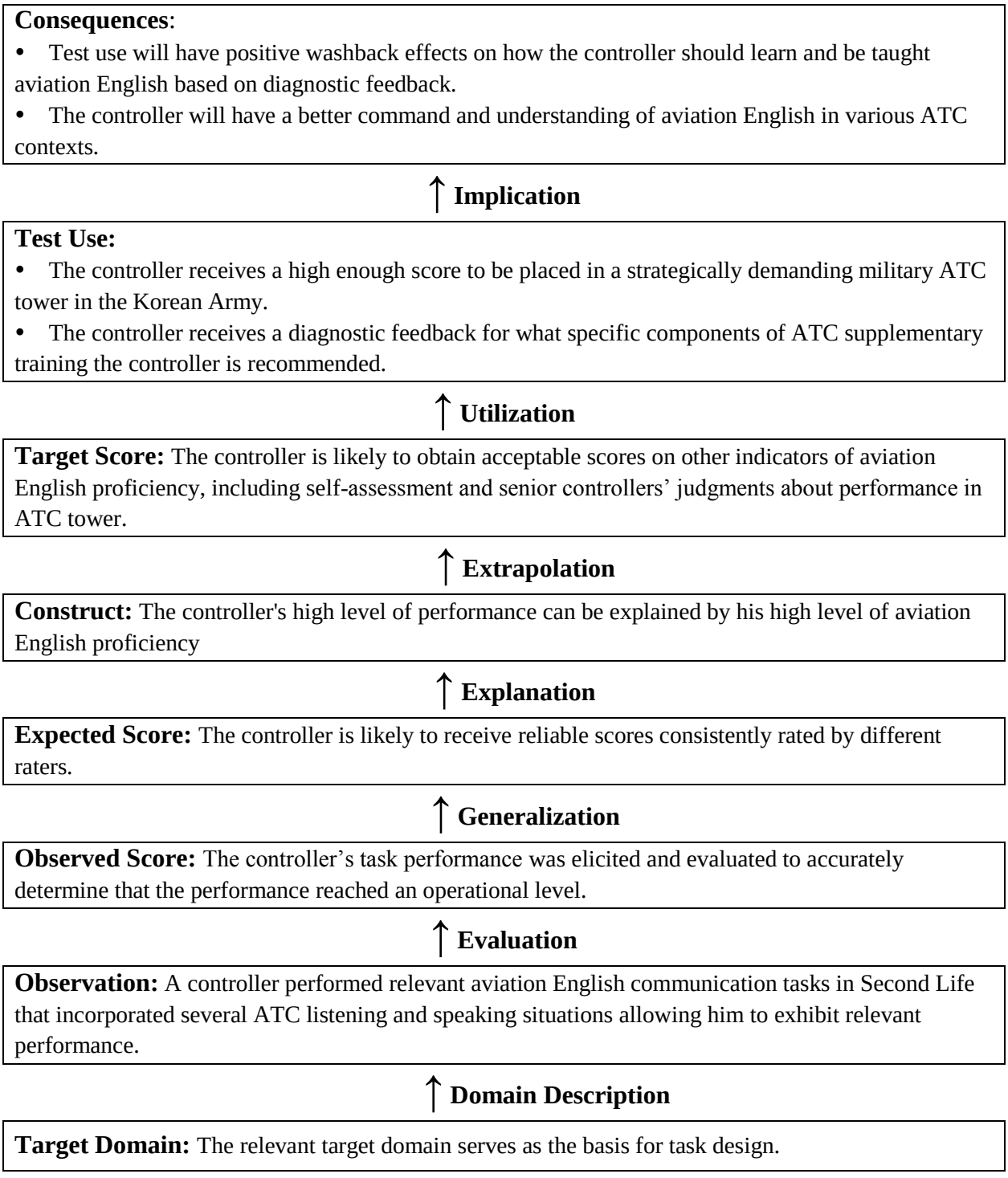

Figure 6. An illustration of the grounds, claims, and inferences in the interpretive argument for a test taker's performance on the prototype virtual interactive tasks for aviation English assessment

As a test developer, the researcher stated the final conclusion in the interpretive argument as the consequences of using the prototype virtual interactive tasks for aviation English assessment. By using the virtual interactive tasks for aviation English assessment, it is expected that the test use will have positive washback effects on how the controller should learn and be taught aviation English based on diagnostic feedback. Additionally, the controller will have a better command and understanding of aviation English in various ATC contexts. Intended test use that could be made on the basis of the assessment-based interpretation involves the usefulness of the virtual interactive tasks for aviation English assessment scores in making decisions about whether the controller is capable enough to be placed in much busier and more demanding air traffic control towers and what specific components of ATC supplementary training the test taker is required to receive. The intended decision is based on the interpretations about air traffic controllers' scores from the performance in the virtual interactive tasks for aviation English assessment, as the score reflects test takers' level of ATC proficiency and ATC task accomplishment. The interpretation about the controller's language ability is based on the observed score of the controller's task performance and the expected score is the intended interpretation of the observed score as an indicator of the score that the controller would be likely to receive in other aviation English tasks included in the defined universe of relevant tasks. The first claim that is made in the interpretive argument is about the target domain, which is taking the prototype aviation English test in a virtual world which requires air traffic control knowledge and aviation English communication ability.

To complete the outline of the army aviation English test interpretive argument shown in Figure 6, the specific warrants and assumptions associated with each of the inferences must be identified. The proposed warrants and assumptions entailed in the aviation English test

interpretive argument are shown in Table 23, which also provides an illustration of the type of analysis that can be used to provide backing for each assumption. The following sections outline the inferences, each containing a supporting warrant, assumptions underlying the warrant, examples of backing for the assumption, and a rebuttal. All of which are associated with the interpretive argument for the virtual interactive tasks for aviation English assessment. As the first prototype phase of a new test development, the current dissertation research investigated the first four inferences in the interpretive argument framework – domain description, evaluation, generalization, and explanation.

Table 23.

Summary of the Warrants, Assumptions, and Backing in the Interpretive Argument

Inference in the Interpretive Argument	**Warrant Supporting the Inference**	**Assumptions Underlying the Warrant**	**Analysis to obtain Backing for Assumptions**
Domain Description (Target language use domain → observation of task performance)	Observations of performance on the virtual interactive tasks for aviation English assessment are representative of relevant language/topical knowledge of ATC, and English listening and speaking skills that are necessary for army air traffic control context in Korea.	(1) Critical aviation English skills, knowledge, and processes needed for aviation English communication in the military ATC context can be identified.	Domain analysis (expert consensus, document analysis)
		(2) Assessment tasks that are representative of the English for Specific Purposes (ESP) domain on ATC have been identified.	Domain analysis (needs analysis survey, expert consensus)
		(3) Assessment tasks that require important knowledge, skills, and processes for aviation English communication can be simulated in Second Life.	Systematic process of task design and modeling (expert consensus)
Evaluation (Observation of performance → observed score)	Observations of performance on the virtual interactive tasks for aviation English assessment are evaluated to provide observed scores that reflect test takers' targeted language abilities.	(4) Both analytic and holistic rubrics for scoring performance responses are appropriate for providing evidence of targeted language abilities.	Systematic rubric development
		(5) Task administration conditions are appropriate for providing evidence of targeted aviation English ability	Trial and revision of task administration conditions
		(6) Raters can be trained to avoid rater bias.	Rater training and calibration

Table 23. (continued)

Generalization (Observed score → expected / universe score)	Observed scores are estimates of expected scores over the relevant parallel versions of tasks and within and across raters.	(7) Under ECD framework, task and rating specifications are well defined for parallel task creation.	Systematic task and rating specification development (expert consensus)
		(8) Ratings of different raters are consistent.	Inter-rater reliability
Explanation (Expected/universe score → construct)	Expected scores can be attributed to a construct of aviation English proficiency and integrated abilities for ATC.	(9) Performance in the virtual interactive tasks for aviation English assessment relates to performance on other aviation English assessment (e.g., Pearson's Versant Aviation English Test).	Concurrent correlational studies
		(10) Strategies engaged by tasks are construct relevant and in accord with theoretical expectation.	Discourse analysis of test takers' think-aloud protocol; correlations between strategy use and test scores
Extrapolation (Construct → target score)	The construct of aviation English knowledge and communication skills as assessed by the virtual interactive tasks for aviation English assessment accounts for the quality of ATC performance in Korean army aviation context.	Performance in the virtual interactive tasks for aviation English assessment is related to other criteria of aviation English communication skills and ATC knowledge in the army aviation English communication context.	Criterion-related validity studies (Examination of relationships between test performance and test takers' self-assessment of their own aviation English)
Utilization (Target score → test use/decision)	Estimates of the quality of task performance in the virtual interactive tasks for aviation English assessment are useful for making decisions about controllers' next job placement and follow-up supplementary ATC training.	The test scores provide useful information to ATC training officers and test takers regarding test takers' aviation English communication abilities through a clear understanding of the meaning of the test scores.	Score descriptors are provided to test takers along with their test result; test takers' and ATC training officers' perceived interpretability and usefulness of the descriptors.
Implication / Consequence (Test use → consequences)	The consequences of using the virtual interactive tasks for aviation English assessment and the decisions that are made are beneficial to the controllers and test users.	The test will have a positive influence (washback effect) on how aviation English is learned and taught.	Washback studies (Expert interviews, follow-up controller questionnaires and interviews)

Domain Description Inference. In the Korean Army Aviation English Test's interpretive argument, the first inference, the domain description inference, is based on the

warrant that observations of performance on the army aviation English tests are representative of relevant language/topical knowledge of ATC and English listening and speaking skills that are necessary for the army air traffic control context in Korea. This warrant, in turn, is based on two assumptions: (1) critical aviation English skills, knowledge, and processes needed for aviation English communication in the military ATC context can be identified; (2) assessment tasks that are representative of the English for Specific Purposes (ESP) domain on ATC have been identified; and (3) assessment tasks that require important knowledge and skills for communication in aviation English can be simulated in Second Life. The first two assumptions were supported by backing through domain analysis. Critical knowledge and skills for aviation English communication required in the target domain have been investigated through expert consensus, task-based needs analysis, and textbook analysis, and then compared with the assessment tasks in Second Life, the medium which will simulate the target domain. The potential experts are experienced military air traffic controllers who have served at least 14 years in the ATS Battalion. The aviation English tasks which are being developed in Second Life and its virtual environment will be presented to the ATC experts to elicit their comments and feedback about the appropriateness and affordability of the test tasks and task environment. The document analysis was focusing on an ATC textbook used in the Army Aviation School and an ATC training module manual book (AAS, 2012) used in field training in the ATS Battalion.

Evaluation Inference. The second inference, the evaluation inference, connects the task performance observation with an observed score. The observation of performance in the virtual interactive tasks for aviation English assessment can serve as grounds for evaluation. The evaluation inference is based on the warrant that observations of performance on aviation English tasks are evaluated to provide observed scores that reflect test takers' targeted language

abilities. This warrant is based on three assumptions: (4) both analytic and holistic rubrics for scoring performance responses are appropriate for providing evidence of targeted language abilities; (5) task administration conditions are appropriate for providing evidence of targeted aviation English ability; and (6) raters can be trained to avoid bias. The backing for the first assumption on the test rubrics will come from systemic rating development for a construct-centered and task-centered rating rubric. The rubric will be developed and revised according to the rating criteria recommended by ICAO, expert controllers' consensus, and findings from the task-based needs analysis. The second assumption can be supported with the backing of a task administration condition trail and modification. As the test tasks were implemented in a virtual world, the test administration condition, including the delivery of the text and audio prompts and synchronization of audio- and video-recording systems, were examined. Backing for the third assumption will be realized through rater training and calibration. The fourth assumption can be justified through a warm-up session for test takers before the actual assessment to familiarize them with the virtual environment and use of an avatar.

Generalization Inference. The third inference, which links the observed score to an expected score, is the generalization inference. This inference is supported by the warrant that observed scores are estimates of expected scores across test raters. Generalization can be inferred on the basis of two assumptions: (7) task and rating specifications are well-defined for parallel task creation under the ECD framework and (8) ratings of different raters are consistent. The backing for the first assumption comes from the experts' consensus about systematic task and rating specification development. so that parallel task creation can be ensured (further specified in Chapter 3 section on test development and validation). One of the great strengths of the ECD framework is that the developed test specification is not only a blueprint for test design and

construction as a higher-order organizational tool, but also a means for generating priori validity evidence. Therefore, the results section structured around the generalization inference mainly focuses on backing for rating consistency. The second assumption will be justified through the analysis of inter-rater reliability. Two or more raters did rate the same set of performance assessment responses (recorded audio files) to measure correlation coefficients for the inter-rater reliability analysis.

Explanation Inference. The fourth inference, which links the expected score to the construct, is the explanation inference. This inference includes the warrant that expected scores can be attributed to a construct of aviation English communication ability. There are two assumptions underlying this warrant: (9) performance in the virtual interactive tasks for aviation English assessment relates to performance on other test-based measures of aviation English proficiency; and (10) strategies engaged by tasks are construct relevant and in accord with theoretical expectation. The first assumption can be supported by concurrent correlational studies that examine correlations between the virtual interactive tasks for aviation English assessment and the commercial aviation English test developed by Pearson (the Versant Aviation English Test). The second assumption will be supported by a discourse analysis of test takers' verbal protocols (stimulated recalls) and analysis of screen captured video recorded during test implementation. Additionally, descriptive statistics and data set investigation between strategy use and test scores will be examined.

Extrapolation Inference. The fifth inference, the extrapolation inference, links the construct to the target score. The extrapolation inference is based on the warrant that the expected scores can be attributed to a construct of aviation English communication and integrated abilities for ATC. This warrant is based on the assumption that test taker performance

in the virtual interactive tasks for aviation English assessment is related to other criteria for assessing aviation English communication skills and ATC knowledge in the Korean army aviation context. This assumption can be backed by evidence from backing through an examination of relationships between test performance and test takers' self-assessments of their own aviation English proficiency.

Utilization Inference. The sixth inference, which links the target score to test use, is the utilization inference. Utilization can be inferred based on evidence showing that estimates of the quality of task performance in the virtual interactive tasks for aviation English assessment are useful for making decisions about controllers' next job placement and follow-up supplementary ATC training. The assumption underlying the warrant is that the test scores provide useful information to ATC training officers and test takers in terms of test takers' aviation English communication abilities through a clear understanding of the meaning of the test scores. Backing for this assumption will come from the investigation of test takers' and ATC training officers' perceived interpretability and usefulness of the descriptors.

Implication Inference. The seventh and final inference, which links the test use to the intended consequence, is the implication inference. The warrant that supports this inference is that the consequences of using the virtual interactive tasks for aviation English assessment and the decisions that are made are beneficial to the stakeholders of the test takers (controllers) and test users (ATC training officers) in the ATS Battalion. The assumption underlying the warrant is that the test will have a positive influence (washback effects) on how aviation English is learned and taught in the target context. This assumption will be supported by washback studies through surveys and interviews with the test takers and training officers.

Conclusion. The previous section outlined a series of backing studies needed to support assumptions underlying warrants throughout the interpretive argument, from the domain description inference to the consequence inference. However, this dissertation study will primarily focus on the first three inferences (domain description inference, evaluation inference, and explanation inference) instead of addressing the entirety of the interpretive argument. The development and pilot testing of aviation English tasks in Second Life will serve as the first empirical stage of the development of a new aviation English assessment.

In order to provide evidence to support the backing for the three inferences in the interpretive argument, a prototype of new aviation English assessment tasks was developed in the winter of 2014 and revised in spring of 2015 to collect additional data from the air traffic controllers in the Air Traffic Services Battalion in the Korean Army. The overall validation project has spanned approximately two semesters.

Research Questions

This dissertation study aims to seek empirical evidence to support the warrants underlying the domain description inference, evaluation inference, and generalization inference as the initial prototyping phase of the development of the virtual interactive tasks for aviation English assessment (VITAEA) for Korean air traffic controllers in the Air Traffic Services Battalion. The empirical research findings will serve as backing to support the four inferences (domain description, evaluation, generalization, and explanation). As these four inferences primarily focus on tasks, measurement, and test forms, an in-depth analysis of the beginning four inferences enabled the test developer to modify the test tasks and test environment before full-scale implementation of the test. This prototyping phase includes new assessment task

development as well as the foundation for a validity argument. Through domain description analysis inference, new assessment tasks will be explored to see whether they can be simulated in a virtual environment, eliciting important knowledge and skills of the aviation English domain. Through empirical studies on evaluation inference, the researcher will investigate how each test taker's performance on test tasks could be scored and explain evidence of aviation English proficiency. Through the generalization inference, the reliability of task-based performance assessment rating will be investigated. In the explanation inference, the researcher will seek evidence that the test scores could be attributed to a construct of aviation English proficiency.

Based on the four inferences and required backing identified in the interpretive argument, the following research questions were developed to guide the empirical research that provides backing for domain description, evaluation, generalization, and explanation inferences. The corresponding inference and assumption for each research question is listed as follows.

1. Domain description inference

    1.1. What are the important skills, knowledge, abilities, and processes needed for aviation English communication in the Army Aviation context as identified by expert air traffic controllers and training manual books?

    1.2. What are authentic target tasks that could be representative of the target domain of aviation English communication in the Army Aviation context as identified by the task-based needs analysis conducted with experienced air traffic controllers?

    1.3. What are experts' perceptions of assessment tasks simulated in Second Life?

2. Evaluation inference

    2.1. What are experts' opinions about the use of construct-centered and task-centered rating rubrics for scoring performance responses?

    2.2. What are test takers' perceptions about the task administration conditions for prompting their use of relevant abilities in a virtual environment?

    2.3. To what extent can test raters be trained to avoid rating bias in the task-based performance assessment?

3. Generalization inference

    3.1. To what extent did experts find the test task specification appropriate for producing parallel tasks?

    3.2. How high is the inter-rater reliability?

4. Explanation inference

    4.1. What is the relationship between test performance on the VIAET and Pearson's Versant Aviation English Test?

    4.2. What are test takers' test-taking strategies as identified in a discourse analysis of the think-aloud data?

The following chapter describes the methods used to gather the data used to investigate these questions, and Chapter 5 presents the results. In the final chapter, the researcher concludes by interpreting the results in terms of the extent to which they provide backing in support of their respective parts of the validity argument.

CHAPTER 4

METHODOLOGY

Research Design

This dissertation study adopts a mixed-method research design (Creswell & Plano Clark, 2007) with quantitative data and qualitative data to support the warrants of the four inferences (domain description, evaluation, generalization, and explanation) in the interpretive argument for the aviation English assessment. The central premise of following a mixed-methods research with triangulation design is that the integrated use of quantitative and qualitative approaches holds the ability to provide a better understanding of research problems than if depending on a single approach.

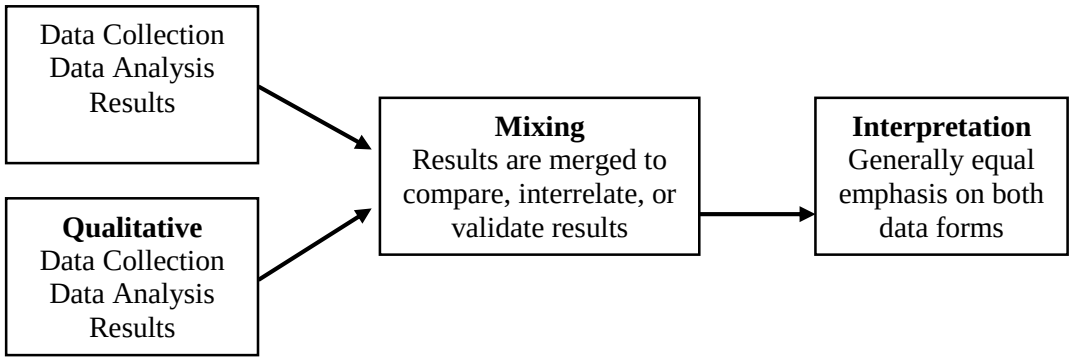

Figure 7. Mixed Method - Triangulation Design (Creswell and Plano Clark, 2007)

As presented in Figure 7, this mixed method with triangulation design includes two primary data sources: 1) quantitative data from closed-item questions in questionnaires, test-takers' aviation English test performance scores, and coded think-aloud data; and 2) qualitative data from open-ended questions in questionnaires, interviews, screen recordings, and raw, think-aloud discourse data. The triangulation design is used to directly compare quantitative and qualitative forms of evidence to corroborate results or identify discrepancies between collected data sources under the interpretive argument framework.

Context and Participants

Context. The context of this dissertation study, where the prototyping of new aviation English assessments in a virtual environment took place, is an Air Traffic Services (ATS) Battalion in the Army Air Operations Command in the Republic of Korea. Due to several restrictions, including security reasons, there has been little to no research into aviation English assessment or analyses of characteristics of military air traffic controller groups in the target context. According to Ryu (2012, 2013), the conventional aviation English test in the Republic of Korea (ROK) Army aviation is not free from issues of inadequate aviation English assessment, construct under-representation, and construct-irrelevant variance.

Over a decade of acknowledging the importance of testing Korean air traffic controllers' aviation English ability, the ATS Battalion in the Army Air Operations Command in Republic of Korea has used the conventional aviation English test, a written multiple choice, grammar, reading, and vocabulary-focused aviation English test that measures novice and experienced air traffic controllers' aviation English proficiency. Though the conventional Korean Army aviation English test has failed to meet the needs of the test users, the primary use of the aviation English test is to evenly distribute novice and experienced air traffic controllers with respect to their demonstrated level of aviation English communication skills. Based on the test results, those who score higher are considered capable of successfully accomplishing particular real-world tasks and being deployed in a more strategically important or busier flight coordination center, ground control center, or airbase tower. Conversely, those who score lower are sent to the ground control center or airbase tower where soldiers' positions and duties are of lesser strategic importance and there is a lighter workload. The secondary use of the assessment, and one in line with the intended purposes of the test, is to provide diagnostic feedback to the air traffic controllers in order to improve their aviation English communication skills.

The client (contact person) for this dissertation study was Command Sergeant Major (CSM) Jewon Ryu, who has been in charge of ATC operational training and testing in the ATS Battalion in the Korean Army for the last seven years. CSM Ryu has excellent aviation English communication proficiency with in-depth knowledge and experience of air operation with U.S. military air traffic controllers, though his English pronunciation is slightly influenced by his first language (L1). He has served in the Korean Army as an air traffic controller for nearly 22 years, and worked at the same ATC center with the researcher from 1998 to 1999.

Participants. Participants for this dissertation study are Korean army enlisted soldiers and noncommissioned officers (NCO) whose job is to conduct air traffic controlling in the ATS Battalion. Access to the research participants was made possible by CSM Ryu. The participants in the study are categorized according to research questions. Participants (test takers, including both enlisted soldiers and NCOs for the virtual interactive tasks for aviation English assessment, were recruited from the military airbase, where the headquarters of the ATS battalion is located, from December, 2014 to January, 2015 with the aid of CSM Ryu. In fact, the expected voluntary participants were from the participant group who already joined the task-based needs analysis, reported in the task development section. Among the 81 task-based needs analysis survey participants, 20 military air traffic controllers agreed to participate in this dissertation study by signing the informed consent document before beginning the aviation English assessment. The 20 participating military air traffic controllers (16 males and four females) consisted of eight enlisted soldiers and 12 NCOs from the ATS Battalion. All of the four female participants are NCOs. A brief summary of the participants is provided in Table 24.

Table 24

Descriptive Statistics of the Participants' Demographic Information

Measure	Mean	SD	Min	Max
Age	27.00	6.10	22	46
- Enlisted soldiers (*n* = 12)	24.13	1.64	22	27
- NCO (*n* = 8)	28.92	7.26	23	46
ATC service period (in month)	38.60	71.24	0	276
- Enlisted soldiers	5.25	6.39	0	15
- NCO	60.83	85.98	0	276
English test (TOEIC) score	753.80	155.86	530	970
- Enlisted soldiers	858.00	87.50	750	970
- NCO	597.50	76.32	530	700
Study abroad period (in months)	18.00	35.52	0	120
- Enlisted soldiers	45.00	45.13	0	120
- NCO	0	0	0	0

Note. The total number of the participants was 20.

Age. The mean of the responding participants' age, including enlisted soldiers and noncommissioned officers, was 27 and ranged from 22 to 46, which indicates that participating test takers are of a relatively young group in their 20's and 30's. Considering the characteristics of this highly stressful and physically demanding ATC job, this finding is consistent with the job requirements for the target context and other civil and military aviation contexts as well. Descriptive statistics of each group indicate that there is a noticeable difference in ages between the two groups. The average age of the enlisted soldier participants, who are obliged to serve in the military, is 24.13, four years younger than the NCO group on average. The small standard deviation value of the enlisted soldier group's age (*SD* = 1.64; *range* = 22 – 27) and its narrow range indicates that most ATC enlisted soldiers are in their early twenties. The age mean of the NCO group is 28.92, distributed from 23 to 46, with a standard deviation of 7.26. Considering the NCO retirement age of 55, the ATC NCO group in the TLU situation is also a fairly young group.

ATC service period. The mean of both enlisted soldier and NCO participants' ATC service period is 38.60 months, yet its large SD value indicates that there could be a big difference between the two groups. When it comes to the military service time period for enlisted soldiers, they are required to serve for 21 months including the two month time period from the training at the recruit training center and army aviation school. Contrary to enlisted soldiers' short (approximately four weeks) ATC training period at a military aviation school, noncommissioned officers are trained for about 12 weeks at the aviation school. As the enlisted soldier controllers' compulsory service term is 21 months, this finding reflects that the participants are also comprised of noncommissioned officers who have worked much longer periods of time than the enlisted soldiers. The average ATC service period for the enlisted soldier group is 5.25 months (*SD* = 6.39; *range* = 0 – 15), while that for the NCO group is 60.83 months (*SD* = 56.43; *range* = 0 – 276). This ATC service period suggests that the NCO group has acquired more practical ATC field experience and skills through their prolonged period of ATC service.

English test (TOEIC) score. To prove their English communication proficiency, all NCO and enlisted soldier air traffic controllers are recommended to submit a standard English test score, such as Test of English as International Communication (TOEIC), prior to initiating their military service as air traffic controllers. As for enlisted soldiers, their average TOEIC test score is 858.00, and ranged from 750 to 970. On the other hand, the average TOEIC test score of the NCO group is 597.50 with a range from 530 to 700, which is much lower than that of the enlisted soldiers. The higher English proficiency test score of the enlisted soldier group may imply that ATC enlisted soldiers could be more confident and fluent, in general, with colloquial English as compared to the ATC noncommissioned officers.

Study abroad period. The average period of time spent in a study abroad experience in English speaking countries in the ATC enlisted soldier group is 45.00 months (*SD* = 45.13; range = 0 – 120). The analysis of the individuals' responses indicates that all of the enlisted soldiers except one (seven out of eight) ever experienced studying in English speaking countries, such as Canada, Philippines, U.S., and U.K., whereas none of the NCOs ever experienced study abroad in the past. This finding indicates that ATC enlisted soldiers are generally more exposed to general (colloquial) English and the culture of English speaking countries compared to the NCOs.

Academic major. The findings from the participants' bio-data responses show that participants' average education level both in the enlisted soldier group and the NCO group represented attendance at a two-year college at least. Regarding the academic major of participants' higher education, all of the NCOs had to study in the NCO Academy as part of the required higher education in the military. Meanwhile, ATC enlisted soldiers were from diverse higher education majors, including Business (*n* = 1), Economics (*n* = 1), Engineering (*n* = 2), English language (*n* = 2), Law (*n* = 1), Psychology (*n* = 1), and Statistics (*n* = 1).

Military rank. In the military TLU situation, there exists a four-level rank system in the enlisted soldier group (listed from the lowest to highest rank): private second class (PSC) as the lowest rank; private first class (PFC); corporal (CPL); and sergeant (SGT) as the highest rank. The NCO group also adopts a four-level rank system (listed from the lowest to highest rank): staff sergeant (SSG); sergeant first class (SFC); master sergeant (MSG); and sergeant major (SGM). As for enlisted soldiers, it takes three months to progress from PSC to PFC, seven months from PFC to CPL, seven months from CPL to SGT, and four months from SGT to retirement. In the TLU situation, enlisted soldiers with a rank of corporal or higher are generally

believed to be competent air traffic controllers, and they usually work with lower ranked (PSC and PFC) enlisted controllers, assisting them to become more comfortable and confident in ATC communication. Noncommissioned officers take charge of ATC communication, but they are also required to lead the ATC enlisted soldiers and perform administrative duties.

In summary, with regard to the characteristics of the test takers of the current study, ATC enlisted soldiers have more experience with study abroad in English speaking countries and have much higher standard English test scores compared to NCOs. On the other hand, despite lower standard English test scores, ATC noncommissioned officers have had intensive ATC training in the military aviation school, including aviation English communication, and more prolonged experience with ATC communication in a variety of situations. Such extended and diverse ATC experiences could also lead NCOs to be confident and competent in ATC communication.

Materials and Instruments

This section summarizes data collection methods and instruments. The data collection instruments developed for this study include: a task-based needs analysis survey for research questions 1.1, 1.2; the semi-structured interviews for experts in response to research questions 1.3, and 2.1; post-test questionnaire/interview for test takers for research question 2.2; pre-recorded task performance audio files with rated scores for research question 2.3; rating rubrics for research questions 3.1 and 3.2; and Pearson's Versant Aviation English Test for research question 4.1.The corresponding research question for each material is listed in parentheses.

Semi-structured Interviews for Experts and Stakeholders. Semi-structured interview protocols for focus group interviews with expert air traffic controllers and stakeholders in the ATS Battalion were created (see Appendix F). Interviews with expert air traffic controllers and

stakeholders were conducted to identify expert judgment of the TLU domain analysis, focusing on skills, knowledge, and processes, expected aviation English tasks, language-centered and task-centered rating rubrics, task administration conditions, and rating calibration. These interview questionnaires were developed to seek backing to support assumptions in domain description inference and evaluation inference.

Simulated Aviation English Task Environment in Second Life. The virtual interactive tasks for aviation English assessment (VITAEA) entail listening to simulated pilots' radiotelephony transmissions involving requesting information, instructions, and clearance, watching animated helicopters landing and departing, reading flight plans and weather information in a digital board in the ATC tower, and then orally responding to the pilots in seven prototype aviation English tasks simulated in Second Life.

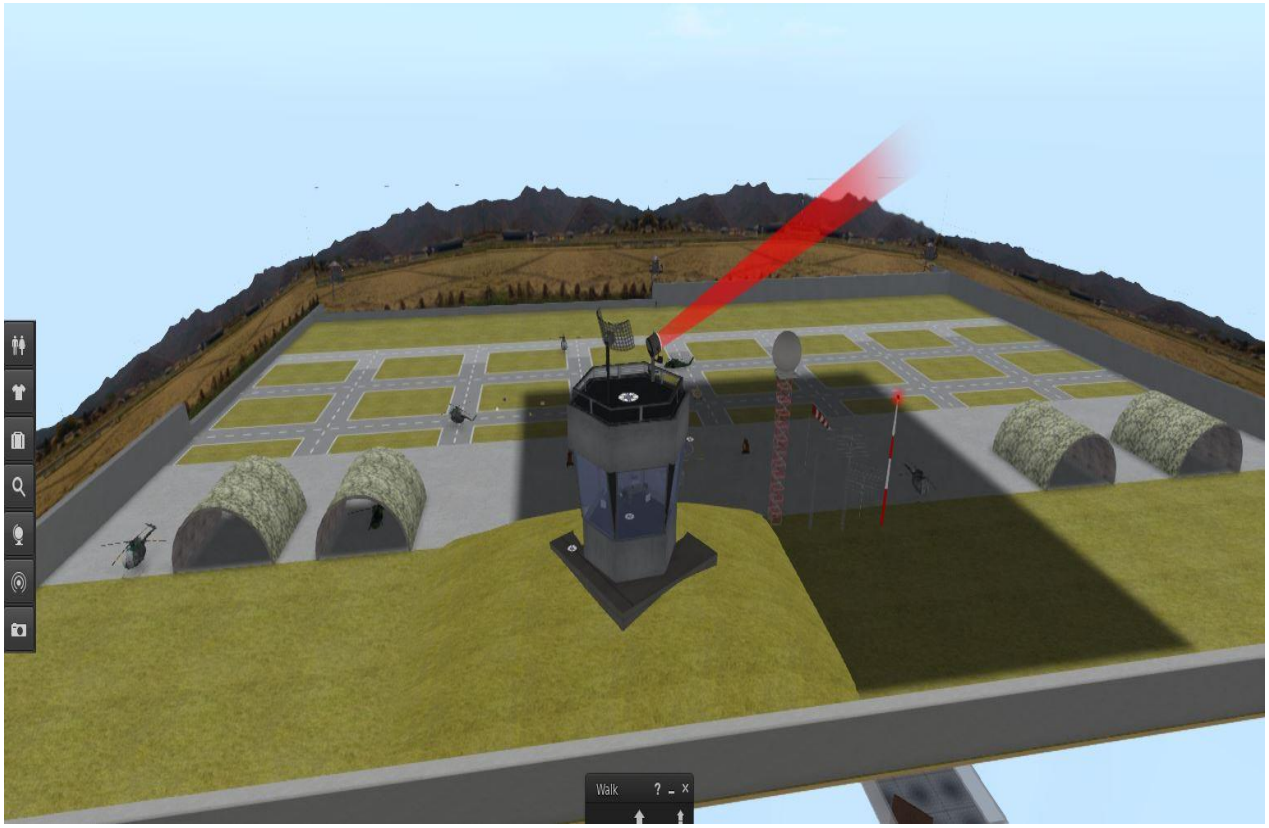

Figure 8. Overview of the VITAEA environment simulated in Second Life

As shown in Figure 8, test takers engaged in aviation English communication in the simulated air base environment which was created based on the authentic satellite images and pictures from the actual air base in the Republic of Korea. The structure and placement of runways, the ATC tower, hangars, windsocks, antennas, helicopters, and a refuel truck were created with Dr. Sadler using CSM Ryu's feedback and confirmation. When test takers log into ATC tower environment in the testing situation, they can interact with simulated pilots and helicopters and access ATC information, such as weather information, air base information, Notice to Airmen (NOTAM), and flight plans. When test takers log into the testing environment, the avatar of the test taker himself or herself is standing in the ATC tower as a default mode.

Virtual Interactive Aviation English Tasks. The analysis of target tasks according to four language skills in the target language use (TLU) situation revealed the characteristics of individual target tasks and how target tasks from one English skill set are interrelated with other skills. Considering the interrelated nature of aviation English tasks in TLU situations, the results from an analysis of task topic across four language skills identified 14 task topics: weather, flight plan, transition, landing, pilot report (PIREP), traffic flow, departure, emergency flight, notice to airmen (NOTAM), layover information, aviation phraseology, flight log, colloquial English, and position report. A language skill analysis showed what specific language skills were involved with the specific topic. Total occurrence presents the portion of individual task topics out of the entire 275 target tasks from the open-ended survey questions. Approximately 19% (n = 52) of target tasks were related to weather, followed by the topics flight plan, transition, landing, PIREP, traffic flow, and departure. The 14 task topics can be further categorized into three primary tasks of landing, departure, and transition, and with other task topics playing roles in supporting those three primary tasks.

Based on this authentic task and task-topic identification process, seven prototype aviation English assessment tasks were created. The prompts for the virtual interactive tasks for aviation English assessment are presented in Appendix E. The first task aims to assess test takers' colloquial English communication in the target context, such as test takers' plans for the weekend. The second task aims to assess test takers' listening comprehension of a changed flight plan. The third task aims to assess test takers' performance of departure information transmission. The fourth task aims to assess test takers' performance of departure procedure transmission. The fifth task aims to assess test takers' performance of transition procedure transmission. The sixth task aims to assess test takers' performance of arrival procedure transmission. Lastly, the seventh task aims to assess test takers' multi-task performance of concurrent ATC situations involving arrival and departure. Test takers' oral responses were automatically saved as audio files for rating. Test takers' task performance in Second Life was also video-recorded using screen capture software called *Camtasia Studio* (Studio, 2007).

Language-Centered Rating Rubric. The evaluation of test takers' aviation English performance in Second Life was based on two different rating criteria: (1) an ICAO-implemented language-centered rating rubric, and (2) a task-centered rating rubric focusing on the task accomplishment. In the Manual on the Implementation of ICAO Language Proficiency Requirements (2004), six levels of operational proficiency (see Appendix C), ranging from pre-elementary (Operational Level 1) through expert (Operational Level 6), are provided. There are six dimensions of proficiency that are evaluated:

- **Pronunciation** - pronunciation, stress, rhythm, and intonation,
- **Structure** - grammar, sentence patterns, global meaning errors, and local errors,

- **Vocabulary** - style, tone, lexical choices which correspond to context and status, idiomatic expressions, articulation of subtle differences or distinction in expression, and meaning,
- **Fluency** - naturalness of speech production, absence of inappropriate hesitations, fillers, and pauses that may interfere with comprehension,
- **Comprehension** - clear and accurate information transfer that results in understanding), and
- **Interactions** - sensitivity to verbal and non-verbal cues and appropriate response to them

Test takers' performance was rated based on the six constructs of aviation English proficiency (see Appendix C).

Task-Centered Rating Rubric. To provide diagnostic feedback to stakeholders and test takers, test takers' aviation English performance in Second Life was rated based on the task-centered rating criteria. A primary concern of the task-centered rating rubric is test takers' accomplishment of target tasks in TLU situations. Raters will be asked to identify the level of task accomplishment as (1) excellent, (2) acceptable, or (3) unacceptable levels of task accomplishment. Task-centered rating criteria in each level correspond to one another (see Appendix D). The prototype of the task-centered rating rubric was revised according to experts' consensus and feedback before piloting.

Pearson's Versant Aviation English Test (VAET). To justify an explanation inference assumption where assessment records in the current virtual interactive tasks for aviation English assessment are consistent across different forms of aviation English assessment, Pearson's Versant Aviation English Test was adopted for concurrent correlation studies. The Versant Aviation English Test utilizes speech recognition technology and a computerized testing platform so that test administration and scoring are fully automated. Developed in collaboration

with the U.S. Federal Aviation Administration, this 25-minute test is delivered via a telephone or computer (Van Moere et al., 2011). A description of the tasks in the VAET is provided in Table 25 (for more information, see Pearson, 2008).

Table 25

Tasks, Prompts and Expected Responses in the VAET

Task	Explanation
A. Aviation Reading	Read aloud sentences that are printed on the test paper or screen, one at a time, in the order requested.
B. Common English Reading	Prompt: 'Now read sentence one'. Response: 'The flight was delayed because some technical issues needed to be resolved.'
C. Repeat	Listen to a sentence and repeat the sentence aloud word-for-word. Prompt: 'I usually get there a couple of hours before the scheduled departure time'. Response: 'I usually get there a couple of hours before the scheduled departure time.'
D. Short Answer Questions	Listen to a question and provide an answer with a word or phrase. Prompt: 'What is the name for land that is surrounded by water on three sides?' Response: "A peninsula.'
E. Readback	Listen to a radiotelephony message and give an appropriate readback Prompt: 'World Air 395, maintain flight level 070.' Response: 'Maintaining flight level 070, World Air 395.'
F. Corrections & Confirmations	Listen to a radiotelephony message and read the call sign on the test paper or screen. The message might contain correct or incorrect information, or a request for information. Respond appropriately. Prompt 1: 'World Air 395, continue descent to flight level 110, report passing 150' Prompt 2: 'Descending to flight level 10 thousand, report passing 15 thousand, World Air 395.' Response: 'World Air 395, negative, continue descent to flight level 110, report passing 150.'
G. Story Retelling	Listen to an aviation scenario and then describe what happened in your own words.
H. Open Questions	Listen to spoken questions and describe your own opinion or experience.

For this concurrent correlation study of this dissertation research, Dr. Alistair Van Moere, the Vice President of Product and Test Development at Pearson, provided 30 Versant Aviation English Test online accounts for free.

Procedure

Test Administration, Verbal Protocol, and Post-Test Interviews. The virtual interactive tasks for the aviation English assessment were administered with the 20 military air traffic controllers one at a time at the bachelor enlisted quarter (BEQ) building in the airbase in the Republic of Korea from February, 2015 to March, 2015. There were three laptop computers used for the test administration. Test takers were seated at a Dell Alienware laptop computer (the first one), one of the best high-performance PC gaming laptops on the market, and the laptop computer had been set up with a screen capturing program called Camtasia to record test takers' performance during the test. On the screen, test takers could see the simulated ATC tower environment in the Second Life program via wireless Internet connection.

There were two additional laptop computers next to the test takers' computer. The second laptop was used for the simulation test administration in Second Life by the researcher. This second laptop was also connected to the Second Life testing environment and the researcher operated helicopters' movement while playing audio files of the pilots in the simulated airbase environment. The third laptop computer was connected to the same Second Life sever, and an invisible avatar was placed at the ATC tower right behind the test takers. The role of this invisible avatar was to provide a high quality screen capture angle for recording video on how the test taker avatar behaved (e.g., changing avatar's viewpoint, moving mouse pointer, making the avatar stand up, etc.) during the task performance. In addition to the three laptops, a video camera was placed behind the test taker to record the entire scene of test performance. Lastly, a digital MP3 voice recorder was placed next to the laptop for test takers to record their oral responses during the task performance and the following stimulated recall data.

Due to this complicated recording set up and the time restriction of the actual air traffic controllers due to their many daily duties, the test was implemented one test taker at a time. The time duration of the test taking itself was about 16 to 20 minutes per person. When combined with follow-up stimulated recall and a post-test interview, it usually took about one hour per each test taker. Accordingly, it was extremely challenging to collect data with more than four participants per day.

When test takers were seated, the researcher explained what this assessment was for and their rights as voluntary participants in this research experiment. When they agreed to participate, participants were asked to sign the informed consent form. When this paperwork was completed, test takers could finally see the simulated testing environment in Second Life. As none of the participants had ever experienced Second Life prior to this project, the test takers were given approximately ten or more minutes to explore the simulation environment, walking in and out of the ATC tower, flying over the airbase, touching the helicopters on the ramp, and so on in the virtual airbase.

When the test takers were ready for the actual test, all of the synchronous recordings in various devices were started as well. The task performance session lasted about twenty minutes, but some participants with limited or no experience of ATC sometimes finished three to five minutes earlier than the other test takers. Right after the test takers finished the test, the researcher saved the screen recording in the Camtasia software in MP4 format on the external hard drive for the retrospective verbal protocol (stimulated recall). Due to the large file size of the screen capture video, it usually took about five minutes to complete the file saving.

While the screen recording was saving, the test takers were introduced to the procedures involved in a retrospective verbal protocol (stimulated recall) with examples of verbal reporting.

This stimulated recall (SR) method involves video-taping test takers' performance during the test then playing back to test takers the recorded video of their own task performance. To counter theoretical and practical shortcomings of the stimulated recall method (Gass & Mackey, 2000; O'Brien, 1993), "Dry-run" (O'Brien, 1993, p. 217) was implemented to help the test takers become familiar with the SR method before the beginning of actual SR data collection with their own task performance video. During the dry-run, the test takers were introduced to the stimulated recall method and processes. After the introduction provided test takers with the airbase information, participants were asked to respond to an audio-recorded pilot's request for departure. Then, the researcher interviewed the test takers using the following sample questions:

- What were you thinking when you decided to say (do) that?
- What were your thoughts of saying (doing) this?
- Why did you decide to do (say) that?

In some cases, test takers tended to present themselves more favorably and/or defensively by creating explanations, and the tendency could be a threat to validity. To prevent such distortion of test takers' recall data, the researcher emphasized that the purpose of this data collection is mainly to understanding test takers' cognitive process (why they chose to say or act in certain ways) and reassured anonymity of the collected data.

After the SR data collection training, the researcher played back the actual screen captured video of test takers' performance; the display showed the face and upper body of the test takers with synchronous movement of the Second Life testing environment including their avatar with recorded audio sound. Test takers were encouraged to make verbal utterances after each radio communication with the pilots. If they forgot to comment, the researcher asked the test takers some of the sample questions introduced earlier. This SR data collection lasted about

30-35 minutes. When the SR interview was complete, the test takers were invited to participate in the post-test interview, which was semi-structured to identify their impressions about the prototype virtual interactive tasks for aviation English assessment with its limitations and potential. This post-test interview took about 10-25 minutes and it was also recorded using a digital MP3 voice recorder.

Test Rating. Two expert (Master Sergeant) air traffic controllers, MSG Kim and MSG Cheong, with more than 15 years ATC experience each were recruited to rate the test takers' task performance using a task-based performance rating rubric (see Appendix D). Three Ph.D. graduate students in the U.S. (two nonnative and one native English speaker) were also recruited to rate the same task performance audio files, but using ICAO's Language Proficiency Requirements Rating Rubric (see Appendix C).

Rater Training and Calibration. Research question 2.3 aimed to investigate to what extent test raters could be trained to avoid rating bias in the task-based performance assessment. Despite the advantage of task-based assessment, the reliability of ratings is one of the major issues in performance-based assessment. To answer the research question, a post-rating semi-structured interview and online survey were administered to explore the quality of the rater training session, raters' understanding and the use of the rating criteria, difficulties during rating, and suggestions for changes. Both rater groups, task-centered and language-centered, were invited to participate in the post-rating interview and online survey. Then, all of the three raters who rated with the language-centered rating rubric responded anonymously to the online survey, but only one rater who rated with the task-centered rating rubric responded with a short text message excusing her/himself with an emergency stand-by due to a North Korean threat near the 38^{th} parallel in Korean peninsula.

The task-centered rater, the experienced air traffic controller, commented that the initial rater calibration session focusing on three sequential steps (familiarization, norming, and practice) with two test takers' performance data helped the rater adjust to the rating criteria. As the rating scale includes only three levels – excellent, acceptable, and unacceptable – and the two expert air traffic controllers have served in the test task situation for over 16 years, they seem to mutually share a clear distinction of such a mastery level based on their experience.

Online open-ended survey responses from the three applied linguists who rated with the language-centered rubric were analyzed in the following order: quality of rater training session; understanding the rating criteria; difficulties raters experienced; and suggestions for improvement. As for the quality of the rater training session, all of the three raters agreed that the calibration session was adequate and the descriptors in the rubrics were clarified and elaborated. However, one of the raters pointed out that rating calibration with two sample performances was not enough to gain confidence in effective rating. Regarding raters' understanding and the use of the rating criteria, raters mentioned that their comfort level using the ICAO criteria increased as they proceeded with the rating. Overall, findings from the analysis revealed that the three raters seem to agree in the rating results that they came up with for the task performances that they rated. Regarding the rating process and results, the three raters commented as follows:

> I used the ICAO heavily during my rating, having the rubric open during the duration of listening to each recording. My comfort level using the ICAO criteria increased as I proceeded with the rating, so after I had completed the rating, I returned to some of the earlier rated Mp3 files to make sure I had provided the test taker with an appropriate score. The criteria were detailed and emphasized the functionality of the speech (how the speaker would/could perform in real world Aviation contexts). These criteria were crucial to my rating; it enabled me to remain consistent with how I rated, as I evaluated each test taker using a standard set of guidelines. (Rater A)

> I believe the quality of the rater training session was adequate. I gained enough practice before I began rating to know how to accomplish the rating. (Rater B)

> During the rater training session, the descriptors in the rubrics were clarified and elaborated. (Rater C)

To sum up, three types of backing discussed above (i.e., appropriate scoring rubrics, task administration conditions, and rater training and calibration) could serve as validity evidence to support the assumptions underlying the evaluation inference shown in Table 2. Such an in-depth and extensive approach to investigating the backing leads to the claim that the observed aviation English test task performances in a virtual environment were accurately assessed, thereby providing the foundation necessary to continue to the generalization inference.

Pearson's Versant Aviation English Test. In the summer of 2015, all 20 test takers who had completed the virtual interactive tasks for the aviation English assessment were contacted again by the researcher (via email) and by CSM Ryu (in person) and invited to take Pearson's Versant Aviation English Test. Fifteen participants responded and took the Pearson's aviation English test using their personal laptop computers. Each test lasted approximately 35 minutes and the test scores were released soon after the completion of the test owing to automatic speech recognition (ASR) technology-based automated scoring.

The results were automatically saved on the testing webserver of the Pearson testing company, and the researcher was able to check the individual test scores and feedback online. Individual test results were sent to each test taker for their reference.

Follow-Up Online Questionnaire and Interview. Those 15 participants who completed both virtual interactive tasks, the aviation English assessment and Pearson's Versant Aviation English Test, were invited to complete a follow-up online questionnaire via email or in person.

Data Analysis

Prior to the main analysis of this dissertation study, a preliminary step to grasp the target

language use situation focusing on the development of a target task and rating rubric was explored, the findings of which are reported in Chapter 3 on Task Development and Validation. In this section, a description of the data analysis in response to each research question is provided.

1.1: What are the important skills, knowledge, abilities, and processes needed for aviation English communication in the Army Aviation context as identified by expert air traffic controllers and training manual books?

Air Traffic Control training manual (AAS, 2012) and aviation English reference books (ATC training manual created by army aviation, ATC training manual created by air force, and situational ATC procedures created by army aviation) adopted by the Army Aviation School and the ATS Battalion in the target context were referred to in identifying important knowledge and processes needed for aviation English communication. Due to confidentiality of the documents, only limited information from the findings was allowed to be reported in this dissertation study. In addition to the findings from the document analysis, open-ended responses from the task-based needs analysis survey about the expert air traffic controllers' perceived skills, knowledge, abilities, and processes needed for aviation English communication in the target context were also examined and reported.

1.2: What are authentic target tasks that could be representative of the target domain of aviation English communication in the Army Aviation context as identified by the task-based needs analysis conducted with experienced air traffic controllers?

Identification of authentic target tasks representing the TLU situation was the fundamental goal of the test development. The data analysis and the process of developing authentic target tasks were detailed in Chapter 3 on Test Development and Validation.

1.3: What are experts' perceptions of assessment tasks simulated in Second Life?

After the pilot version of the virtual interactive tasks for aviation English assessment was developed, two domain experts (CSM Ryu and MSG Kim) were invited to experience and provide feedback on the target tasks and the virtual task environment in Second Life. Due to security restrictions, a large portion of their feedback was not audio-recorded, but instead summarized in the form of written memo. Both the memo and transcribed feedback were analyzed and categorized according to the content, interface, and recommendations for revision.

2.1: What are experts' opinions about the use of construct-centered and task-centered rating rubrics for scoring performance responses?

Three expert air traffic controllers' (CSM Ryu, MSG Lee, and MSG Kim) opinions about the use of language-centered and task-centered rating rubrics were investigated prior to the implementation of the virtual interactive tasks in Second Life. For logistical reasons, their opinions about the two rating rubrics, especially those focusing on potential advantages and disadvantages, were examined through phone interviews. Their audio responses were recorded and transcribed, and a content analysis was conducted to identify any quotes that were relevant to this research question.

2.2: What are test takers' perceptions about the task administration conditions for prompting their use of relevant abilities in a virtual environment?

Post-test semi-structured interviews with 20 test takers were used to learn about the participants' perceptions of various aspects of the simulated aviation English tasks based on the questionnaire in Appendix G. Interview responses were recorded using an MP3 digital recorder and then were transcribed. Next, a content analysis was conducted to examine the test takers' perceptions about the authenticity of virtual aviation English tasks, efficiency of the test environment, comparison with paper-based tests, satisfaction in the tasks, interface of the tasks, and suggestions for revision.

2.3: To what extent can test raters be trained to avoid rating bias in the task-based performance assessment?

After the rating, a post-rating semi-structured interview was conducted with each rater (TLU expert) who rated with the task-centered rating rubric; each took approximately 20 to 30 minutes. Raters who rated with the language-centered rating rubric were invited to participate in an online survey. Questions in the online survey were structured to explore the quality of the rater training session, raters' understanding and the use of the rating criteria, difficulties during rating, and suggestions for changes.

3.1: To what extent did experts find the test task specification appropriate for producing parallel tasks?

As illustrated in Chapter 3, the test development of the current dissertation study was based on the ECD framework (Mislevy, Steinberg, & Almond, 2003) to make explicit how rich

assessment data in the VIAET should be established through iterative cycles of analysis, design, development, implementation, and evaluation instructional design decisions for producing parallel tasks. Under the framework for developing target tasks, task-centered rating criteria (see Table 17), a task shell (see Table 19), and a final blueprint (see Table 20) were created. These elements were then evaluated and approved by TLU domain experts, including CSM Ryu, for parallel test task development.

3.2: How high is the inter-rater reliability?

In general, researchers use inter-rater reliability as a generic term for rater consistency between raters in the ordering or relative standing of performance ratings, regardless of the absolute value of each rater's rated score. On the other hand, inter-rater agreement is the degree to which two or more raters using the same rating scale give the same rating score to an identical observable situation (e.g., aviation English task performance in the simulation) (Graham, Milanowski, & Miller, 2012). Accordingly, inter-rater agreement is a measurement of the consistency between the absolute value of raters' ratings, while inter-rater reliability mainly concerns the relative similarity between two or more sets of ratings. For the current dissertation study, the researcher adopted two rating rubrics to assess 20 participants' task performance. Two raters, expert air traffic controllers in the TLU situation, rated the test takers using task-centered holistic scores, ranging from zero (unacceptable level), one (acceptable level), to two (excellent level) (see Appendix D). Additionally, three raters, two male doctoral students (non-native English speakers) and one female postdoctoral researcher (native English speaker) in the discipline of Applied Linguistics at a state university in the U.S. rated the test takers using language-centered holistic scores, ranging from one to six based on ICAO's LPRs (see Appendix

C). For inter-rater agreement of task-centered rating by the two domain expert raters, Cohen's kappa was calculated, and for language-centered rating by the three applied linguist raters, Intra-Class Correlation was calculated with a statistical software, *IBM SPSS Statistics 19*.

To calculate inter-rater reliability of the two raters of the task-centered rating, Cohen's kappa is adopted according to the five assumptions (Altman, 1999; Landis & Koch, 1977).

- Assumption 1: The judgement that is made by the two raters is measured on either an ordinal or nominal variable and the categories need to be mutually excluded.
- Assumption 2: The rated data are paired observations of the same phenomenon.
- Assumption 3: Each response variable must have the same number of categories.
- Assumption 4: Two raters are independent.
- Assumption 5: Two raters are specifically selected to take part in the study.

All five assumptions are met in the context of the task-centered rating.

To calculate inter-rater reliability of the three raters of the language-centered rating, two-way mixed Intraclass Correlation (ICC) is adopted as (1) the same raters rated all test takers and (2) the three raters are the population of raters, not a sample of raters. When same subjects and raters under similar conditions, two-way mixed effects model seems to be the most appropriate reliability measure according to the nomenclature of Shrout and Fleiss (1979). Using a two-way mixed average-measure ICC, inter-rater reliability was assessed to measure the degree that the raters provided consistency in their ratings of language-centered rating across test takers.

4.1. What is the relationship between test performance on the VITAEA and Pearson's Versant Aviation English Test?

The concurrent validity related to the explanation inference in the current validity framework was explored through the correlation coefficient between the two sets of measurements obtained in Pearson's Versant Aviation English Test and the virtual interactive tasks for aviation English assessment (VITAEA) for the same target population (n = 17). As the collected data were non-parametric with rank order score without normal distribution, Spearman's correlation statistic was adopted to compare the paired test scores from the two tests. To run the Spearman's correlation statistic, *IBM SPSS Statistics 19* was used (Turner, 2014). And we will need descriptive statistics for the three variables, too.

4.2. What are test takers' test-taking strategies as identified in a discourse analysis of the think-aloud data?

This research question aims to examine verbal reports from the stimulated recall (retrospective verbal protocol) from the 20 air traffic controllers in addressing two specific research questions: (a) What types of aviation English communication strategies were used in the virtual aviation English task performance, (b) What was the interplay and sequence of these strategies. The participants' verbal responses in the stimulated recalls were transcribed and coded for further analysis.

The primary data from this stimulated recall were the test takers' (n = 20) stimulated recall verbal reports recorded in a digital MP3 recorder, from which inferences could be made about what and how strategies were planned, used, and assessed during aviation English assessment in Second Life. Possible concerns about verbal reporting were considered based on relevant research studies (Afflerbach, 2000; Cho, 2014).

Immediately after administering the VITAEA test in Second Life, the researcher trained the participants in how to verbalize their cognitive / metacognitive process during their task performance. In each radio telephony transmission by the test taker, the participant was encouraged to think out loud whenever he or she clicked on navigation buttons, moved the mouse curser, or read particular information. Informed by previous think-aloud studies (Afflerbach & Cho, 2009; Cho, 2014), the entire corpus data from the transcribed stimulated recalls were the focus of the data analysis so as to identify a variety of strategies test takers might use during the test. Initially referred to by Purpura (1999) and Dörnyei & Scott (1997), the transcribed protocols were analyzed using grounded theory analysis techniques (Corbin & Strauss, 2008) for open coding to identify an array of strategic actions, and then axial coding to group those actions into a certain number of core strategy types for accomplishing aviation English tasks in Second Life (see Table 26). For communication strategy use analysis, a coding scheme was adopted from Dörnyei & Scott (1997) (see Appendix H).

Table 26

The Coding Scheme Emerging from the Course of Data Analysis

Strategy Type	Description and Example	Determination
Perceiving	Strategic actions to identify radio transmission from pilots and current condition of ATC tower, runway, and/or its control zone. Strategic actions include clarifying received information and/or transmission (CLAR) and assessing the situation prior to events <ASSIT>.	To code a strategy into *Perceiving* requires certain types of clarifying actions (e.g., requesting repetition of parts not received) and strategic moves (e.g., moving to another direction at the simulated ATC tower).
Processing	Strategic actions to construct meaning from perceived information for responses. Strategic actions include simply applying rules (APR), linking with prior knowledge for more complex processing (LPK), monitoring thinking process during the task performance <MON>, and using contextual	To code a strategy into *Processing* does not necessarily involve explicit screen behaviors or verbal responses but requires significant involvement of cognitive processing in the form of retrospective verbal responses (e.g., applying experience and/or knowledge into problem solving, and making inference to

Table 26. (continued)

	information to interpret radio transmission (INF)	interpret meaning, supplying missing information, or making decisions)
Evaluating	Strategic actions to appraise or assess one's own <SE>, another's (pilot's) performance <EO>, and the situation itself <ES> after engaging in the task.	To code a strategy *Evaluating* requires any sort of evaluative judgment in stimulated recall responses.

After the initial coding scheme was developed, the researcher (1st coder) consulted an expert verbal protocol analysis researcher (Dr. Byeong-Young Cho) about the coding scheme and revised accordingly. Then, the researcher and a second coder, a non-native English speaking researcher with a doctoral degree in Applied Linguistics coded test takers' strategy use. To assess what degree a set of test takers' transcripts were consistently coded by the two different coders, intercoder agreement for an entire set of data was calculated (Paltridge & Phakiti, 2015).

Summary and Mapping of Research Questions, Data Collection, and Data Analysis

Data collection methods and data analysis methods corresponding to each inference and backing in the interpretive argument are summarized in Table 27.

Table 27.

Summary of Backing, Research Questions, Data Collection, and Data Analysis

	Backing Sought to Support Assumption	Research Question	Data Collection	Data Analysis
Domain Description	Domain analysis (expert consensus, document analysis)	1.1	Open-ended survey; ATC training manuals	Analysis of expert interview transcripts; document analysis
	Domain analysis (needs analysis survey, expert consensus)	1.2	Task-based needs analysis survey	Analysis of closed/open-ended survey responses
	Systematic process of task design and modeling (expert consensus)	1.3	Semi-structured interviews for experts	Analysis of expert interview transcripts

Table 27. (continued)

Evaluation	Systematic rubric development	2.1	Semi-structured interviews for experts	Analysis of three experts interview transcripts
	Trial and revision of task administration conditions	2.2	Post-test questionnaire / interview for test takers	Analysis of 20 test takers' interview transcripts
	Rater training and calibration; Test-taker training with a warm-up session	2.3	Post-rating questionnaire / interview for raters	Analysis of survey / verbal responses
Generalization	Systematic development of test spec for producing parallel tasks	3.1	Obtain expert consensus on test spec and parallel tasks through expert interview	Analysis of experts' perspectives
	Inter-rater reliability	3.2	The same set of task responses will be rated by two or more raters	Calculation of inter-rater agreement
Explanation	Concurrent correlational studies	4.1	Test scores from Pearson's Versant Aviation English Test and the prototype aviation English test	Examination of relationships between the prototype task performance and Pearson test
	Investigation of test completion processes using think-aloud protocol and screen capturing	4.2	Stimulated recall after the test	Discourse analysis of retrospective verbal protocol; analysis of process data

This chapter described the context and the characteristics of participants in the study. The materials and instruments were then described. The procedures took in the study then detailed. Data analysis methods corresponding to each research question were then explained. The following chapter introduces the results, following the research questions presented in chapter 3.

CHAPTER 5

RESULTS AND DISCUSSION

This chapter presents and discusses the results identified from the data analysis. Although the entirety of the interpretive argument was presented above, this dissertation study primarily focused on the first four inferences (domain description inference, evaluation inference, generalization inference, and explanation inference). As described in Chapter 3 on test development and validation, the results from this dissertation research serve as support for the assumptions underlying the four inferences in the interpretive argument as the first empirical stage of the development of a new task-based aviation English assessment in a virtual environment. This section is organized according to the listed research questions, which are grouped together under the inferences in the interpretive argument (i.e., Domain Description Inference – Research question 1.1, 1.2, and 1.3; Evaluation Inference – Research question 2.1, 2.2, and 2.3; Generalization Inference – Research question 3.1 and 3.2; and Explanation Inference – Research question 4.1 and 4.2).

For the data analysis, the following data were analyzed: two expert air traffic controllers' interview transcripts; 81 air traffic controllers' open-ended responses in the task-based needs analysis; two expert air traffic controllers' interview transcripts about the prototype tasks and the task-centered rating rubric; 20 task-based performance assessment sample audio files; 19 follow-up test taker interviews and online survey questionnaires; three language-centered raters' post-rating questionnaire responses; two task-centered raters' post-rating interview transcripts; military aviation English training manuals and references; coded transcripts of 12 test takers' stimulated recall and their actual task performance.

Domain Description Inference

The domain description inference is based on the warrant that observations of performance on the VITAEA reveal relevant skills, knowledge, abilities, and processes in the military air traffic control tower in the Army Aviation. Backing was sought to support the assumptions under the warrant through (a) domain analysis to identify critical aviation English skills, knowledge, abilities, and processes, (b) domain analysis to identify representative assessment tasks, and (c) systemic process of task design and modeling to establish test tasks that requires important knowledge, skills, and processes for aviation English communication can be simulated.

Domain Analysis (Skills, knowledge, abilities, and processes)

Research Question 1.1 aimed to identify important skills, knowledge, abilities, and processes which are required for aviation English communication in the TLU situations. To answer this question, documents (Army Air Traffic Control manual and ICAO's manual) and open-ended survey responses from 81 military air traffic controllers were analyzed. The following two portions of this section will present findings from (1) the document analysis and then (2) the open-ended survey responses.

A portion of the findings on the important skills and processes in air traffic control communication in the Army ATC manual and ICAO manual were already provided in the sections of characteristics and construct definition of aviation English (outlined in Chapter 2). The Army Air Traffic Control textbook showed that the main ATC manual (AAS, 2012) defines the goal of the ATC training as follows:

"*ATC training aims to enable air traffic controllers to provide services to army pilots based on the understanding of the standard procedures and operational guidance; pilots' flight situations; and flight control facilities' instructions and flight information."*

Furthermore, the document defines the purpose of air traffic control as follows:

"*The purpose of air traffic control is to maintain the order and immediacy of the air traffic by preventing collision between two aircrafts and between an aircraft and an obstacle in the vicinity of the control zone.*"

Two critical abilities identified from the two excerpts are (1) comprehending various sources of information and knowledge, such as pilots' transmission, instructions from their own or other flight operation facilities, and aviation procedures and rules, and (2) providing relevant and prompt services to pilots for the sake of air safety. The textbook consists of the following specific topics:

- Airspace and flight control facilities
- Flight information
- Visual flight control procedures
- General ATC practice

As presented in the goals of the ATC training detailed in the textbook, the first half of the textbook provides critical background information for the air operation and air traffic control. The second half of the book specifically focuses on actual aviation English communication according to flight phases and flight situations. In addition to the required knowledge, detailed processes about aviation English communication in the textbook are listed as follows:

1. Departure phase

   Initial contact → Checking terminal information prior to departure →

   Departure communication → Leaving control zone

2. En-route phase

   En-route reports, such as flight plan, changed destination, extended ground time, changed estimated time of arrival, pilot report (PIREP), weather forecast

3. Arrival phase

   Requesting arrival and terminal information

Based on the given ATC procedure above, actual aviation communication between a controller and a pilot is listed in the last chapter of the textbook.

Eighty-one military air traffic controllers were asked to identify the important skills, knowledge, abilities, and processes that are required in the air traffic control tower contexts. Thirty-one participants out of 81 responded and the findings are summarized according to required knowledge and skills, and specific processes in aviation English communication.

Regarding the required knowledge and skills in the TLU situations, participant responses revealed the following themes:

- Memorization of aviation English phraseology and non-standard aviation English
- Weather forecast based on given data and the naked eye
- Knowledge about aircraft type, specification, characteristics during flight, air regulations
- Ability to respond to an emergency promptly and properly
- Maintainability of safe distance in the air traffic

From the responses, the researcher was able to identify detailed aviation English communication processes and group them into three TLU contexts: departure procedure, transition procedure, and landing procedure. Among the 81 survey participants, a total of 31 enlisted soldiers (n = 18) and noncommissioned officers (n = 13) described the specific processes of ATC communication in the control tower contexts. As shown in Table 28, the respondents'

experience in ATC tower communication and/or rank seems to affect the description of the ATC processes in the control tower situation. Exemplary responses from three participants (CPL Cho, SFC Seo, and MSG Kim) were selected for analysis, as their responses were more substantial in content than other participants' responses. CPL Cho is an enlisted soldier controller and has served as a controller for 14 months. The other two controllers, SFC Seo and MSG Kim, are noncommissioned officers and have served as controllers for 14 years and eight years, respectively. MSG Kim, the most experienced among the three, described the processes in a more detailed way compared to the other two controllers. Yet, the overall identified processes of departure, transition, and landing from the 31 air traffic controllers are consistent with the three selected participant responses, with varying degrees of including details.

Table 28

Summary of Survey Respondents of ATC Processes

TLU Situations	CPL Cho	SFC Seo	MSG Kim
Departure	Approve a pilot's departure request → Provide traffic and weather information → Approve take off	Approve a pilot's departure request → Provide terminal information (altimeter, departure runway) → Approve taxi to runway → Provide wind information → Approve takeoff	Approve a pilots' request of engine start on the ramp → Provide altimeter information → Approve taxi from the ramp to runway → Provide terminal information → Approve entering the runway for takeoff → Approve take-off
Transition	Receive a pilot's request for transition → Approve transition → Provide traffic and weather information → Receive the pilot's report of transition completion	Receive initial contact from a pilot → Request current position report → Approve transition → Provide traffic information (if necessary) → Request frequency change after transition	Receive initial contact from a pilot before entering the tower's control zone → Request current position report → Approve transition → Provide transition route, weather information, terminal information → Request frequency change after transition
Landing	Receive a pilot's landing request → Provide traffic, weather, and terminal information → Approve landing	Receive a pilot's landing request → Analyze the aircraft's flight information → Assign landing order → Provide wind information → Approve landing	Receive initial contact from a pilot before entering the tower's control zone → Request current position report → Receive landing request → Provide terminal information for landing → Request to enter the traffic → Provide wind and runway direction → Approve landing → Approve taxi from runway to ramp

What follows is a summary of the detailed aviation English communication tasks incorporated in each process that were identified from the document analysis and survey responses:

1. Departure procedure

   Request from a pilot for engine start on a ramp → Provide altimeter information → Approve taxi from the ramp to runway → Provide terminal information → Give clearance to enter the runway for departure → Give permission to take off

2. Transition procedure

   Provide next reporting point (depending on the departure point) or Request from a pilot for overpassing the tower's control zone → Give permission for transition with airspace traffic and weather information → Report from the pilot about the completion of the transition with frequency change request to the next ATC facility → Give permission for frequency change

3. Landing procedure

   Request from a pilot for landing → Provide airspace traffic condition, weather information, and terminal information to the pilot → Give permission for landing → Provide taxi information to the pilot with the ramp information

In summary, identified skills, knowledge, abilities, and processes needed for aviation English communication in the army aviation context, as shown in the analysis of documents and military air traffic controllers' open-ended survey responses, share many commonalities. Particularly in ATC processes, more experienced controllers actually stated very similar ATC procedures, such as the exemplary situational communication between a pilot and an air traffic controller, as those appearing in the textbook chapter on general ATC practice. As the aviation

English phraseologies and specific situations needing highly structured language are clearly identifiable, findings from this first research question correspond to the findings from the second research question and help the researcher develop more context-specific target test tasks under authentic communication processes.

Domain Analysis (Possible Test Tasks)

Research question 1.2 aimed to identify possible test tasks which can be representative of the TLU situations of army aviation. Identified target tasks, according to four language skills and TLU situations and the process of data analysis, are detailed in Chapter 3 on test development and validation. They included 40 (11 listening, 14 speaking, 8 reading, and 7 writing) aviation English tasks and task situations, as well as task-centered rating criteria.

In summary, important skills, knowledge, abilities, and processes needed for aviation English communication in the Korean Army Aviation context were identified from the document analysis and task-based needs analysis survey. Findings from the resources represent TLU situations and test tasks and could be used as authentic materials for aviation English task development.

Systematic Process of Task Design and Modeling

Research question 1.3 examined domain experts' opinions about the target test tasks developed in the Second Life virtual environment. Interview responses recorded soon after a pilot test in the military base were analyzed to answer this research question. The two domain experts were CSM Ryu and MSG Kim, each of who had served in the army aviation as air traffic controllers for at least 16 years. The experts were asked to share their impressions about and

suggestions for revising the initial stage of the target test tasks and the simulated task environment in the virtual world so that the test tasks and the simulated ATC environment could increase authenticity.

First, regarding representative target tasks in the prototype aviation English tasks for aviation English assessment, the two expert controllers suggested the following revisions in Table 29 to make the dialogue between a controller and a pilot more authentic and natural in the five TLU situations (colloquial communication; departure procedure; transition procedure; and arrival and departure procedure) by adding additional ATC communication and background information and deleting non-standard phraseology.

Table 29

Revisions in the Prototype Virtual Interactive Tasks for Aviation English Assessment

TLU Situation	Initial Task Scripts	Revised Task Scripts
#1. Colloquial Communication	ATC 2: Hey, welcome back to Carol tower. I haven't seen you for a while. How have you been? ATC 1: _____ ATC 2: What is your plan for this weekend?	Q 1. How long have you served as an air traffic controller? Q 2. How do you like your air traffic control job? Q3. What is your favorite Korean food? Can you describe it? Q4. Where can I try the food you like around the base?
#4. Departure Procedure	Pilot 1: Carol ground UC 844 hover check completed ready for take-off ATC 1: _____ Pilot 1: UC 844 roger. Carol ground UC 844 holding short. Ready for take-off.	Pilot 1: Carol ground, UC 844 hover check completed, ready for take-off ATC 1: _____ Pilot 1: UC 844 roger. Contact tower Pilot 1: Carol tower, UC 844 holding short. Ready for take-off.
#5. Transition Procedure	ATC 1: _____ Pilot 2: SP 035 Looking for traffic insight. Lower altitude than me. ATC 1: _____ Pilot 2: SP 035 roger	ATC 1: _____ Pilot 2: SP 035 Looking for. Pilot 2: SP 035 traffic insight. Lower altitude than me. ATC 1: _____ Pilot 2: SP 035 roger. Maintain visual separation.
#7. Arrival & Departure Procedure	Pilot 4: Carol tower UC 823 on ramp VFR to Chilgok area request taxi instruction for take-off, over. Pilot 5: Carol tower SP 971 flight of two passing Chilgok area at 23, 2000 encountering turbulence Now 10 miles East of R-523 request landing over.	Pilot 4: Carol ground, UC (Unicorn) 823 on ramp VFR to Chilgok area, request taxi instruction for take-off, over. ATC 1: _____ Pilot 4: UC 823 Roger, runway 02, A then hold short of B. Pilot 5: Carol tower, SP 971 flight of two passing Chilgok area at 2, 2000 encountering turbulence Now 10 miles East of R-523, request landing over.

For the #1 TLU situation, the colloquial communication TLU situation, the domain experts agreed that a casual greeting between U.S. army aviation controllers and Korean army aviation controllers must be the most common TLU situation, as the experts (interviewees) have engaged in numerous combined ATC exercises with U.S. Combat Aviation Brigade stationed in Korea. Yet, they also pointed out that the initial task question "how have you been" is too simple, and worried that the expected responses from the task might be too short to examine the test takers' casual English communication ability. Reflecting their casual conversation with U.S. army controllers, the domain experts came up with four new questions with topics touching on the ATC job and Korean food.

In the of case of TLU situation #4, the departure procedure, even though the initial task script was developed according to the domain experts' advice, CSM Ryu and MSG Kim suggested that the departure procedure needs to include the additional ATC procedure involving communication between a controller and a pilot. Until a recent date, especially in small army aviation airbases, ATC towers were completely in charge of controlling aircrafts' departure procedure, including ground maneuvering, as these ATC towers were not equipped with ground control facilities. However, more recently, army aviation airbases started to provide ground control service to pilots in addition to ATC tower control. This most recent innovation in the TLU settings influenced the revision of task situation #4. The domain experts actually expect a positive washback effect by adding ground control communication between a controller (the test taker) and a pilot.

In TLU situation #5, Pilot 2's transcript "SP 035 Looking for traffic insight. Lower altitude than me" should have been separated into two different transmissions like "SP 035 Looking for" and "SP 035 Traffic insight lower altitude than me." To identify any aircraft within

visual range, the pilot should take additional time to look around and then report the separation condition with the identified aircraft(s) to the controller. To make the task more authentic, the domain experts recommended audio-recording the script again with clear separation in the transmission. Additionally, the domain experts advised inclusion of "Maintain visual separation," expecting the controller (the test taker) could provide this warning to the pilot.

In TLU situation#7, the departure and arrival procedure, the initial task script was designed to situate very complex air traffic with two almost simultaneous initial contacts from the two pilots on the ground. The domain experts mentioned that this complex situation does happen, but very rarely. When pilots and controllers are communicating, all parties tune in to the radio on an identical channel, which indicates that everyone on the same radio frequency could listen to what a controller and a pilot say to one other. That is, when Pilot 4 made an initial call to the tower, Pilot 5 is told not to interrupt the conversation between Pilot 4 and a controller unless it is an emergency.

Furthermore, the two domain experts were asked to provide feedback to enhance the authenticity of the simulated ATC environment. They specified improvement on three issues: (1) the ratio of airbase runway from a point of view at the ATC tower, (2) sky condition, and (3) movement of the aircrafts.

As the two domain experts had been involved in the design of the simulated airbase, they were comfortable with the task environment. However, when they sat in a chair in the simulated ATC tower using their avatar, they pointed out that their view should encompass a greater portion of the airbase runway. They also expected that the environment would be more authentic if the sky condition would change. The default setting of the sky condition was clear, so this condition did not match the task situation and needed to be revised. Lastly, the experts indicated

it would be much more authentic and interactive if the movement of simulated aircrafts in Second Life could be automated in a way that corresponded to test takers' oral responses via automatic speech recognition technology. A week after this pilot test, the researcher visited Dr. Randall Sadler, a dissertation committee member, at the University of Illinois, Urbana-Champaign to revise the simulated task environment in Second Life by focusing on the experts' recommended aspects. Over three days of intensive task revision, and thanks to the timely support from Dr. Sadler, all of the identified issues – changing the ratio of the runway, setting the sky condition, replacing audio files, and incorporating interactive movement of the aircrafts—were able to be revised.

To sum up, the two expert military air traffic controllers experienced the initial model of virtual interactive aviation English tasks and recommended revisions of the task scripts and simulated environments to enhance the authenticity of the assessment tasks and their situations.

Evaluation Inference

The evaluation inference is based on the warrant that observation of performance on the virtual interactive aviation English tasks is evaluated to provide observed scores reflecting the targeted language abilities. Backing was sought through (a) systematic rubric development, to ensure that the scoring rubric would be appropriate for providing evidence of targeted language abilities, (b) trial and revision of task administration, to ensure that the administration conditions are appropriate for providing evidence of targeted language abilities, and (c) rater training and calibration.

Appropriate Scoring Rubrics

Research question 2.1 examined expert air traffic controllers' opinions about the use of construct-centered and task-centered rating rubrics for scoring performance responses with the purpose of providing evidence of aviation English communication abilities. The transcripts of semi-structured interviews with three expert air traffic controllers' (CSM Ryu, SGM Lee and MSG Kim) were analyzed to find answers to the research question. The three domain experts' opinions about the use of language-centered and task-centered rating rubrics were investigated prior to implementation of the virtual interactive tasks in Second Life. For logistical reasons, their opinions about the two rating rubrics, especially those focusing on potential advantages and disadvantages, were examined through phone interviews. Their audio responses were transcribed and a content analysis was conducted to identify any quotes that were relevant to this research question.

Two primary questions during the interview centered on the expected use of the test scores and the experts' perceptions about the language-centered rating rubric [provided by ICAO (2004)] and the task-centered rating rubric (created by the researcher). The domain experts' responses to these questions are summarized in Table 30.

Table 30

Experts Opinion on Language-Centered and Task-Centered Rating Rubrics

	Language-centered rubric	Task-centered rubric
CSM Ryu	• Rating results based on language-centered rubric can be used as a secondary reference to understand test takers' aviation English proficiency more leaning toward colloquial English.	• As conventional air traffic controller evaluation has adopted three-ranking system (Controller level A, B, and C), dividing task performance into three levels look very suitable.

Table 30. (continued)

SGM Lee	• Rating results based on language-centered rubric is needed as the context needs colloquial conversation with U.S. army aviation controllers and pilots.	• If I have to choose from the two, I would use task-centered rubric as the results could directly inform what the test takers could do or not.
MSG Kim	• Language-centered rubric by ICAO's LPR seems to be good lens to identify test-takers' aviation language proficiency, but it does not inform their actual task performance ability in the actual tower situation, which is a shortcoming of this rubric.	• Although there seems to be much room for improvement, this prototype task-centered assessment and rating appear to be optimal for army aviation context.

Regarding the expected test use, the domain experts all expected that the test results could provide diagnostic feedback, which provides certain evidence on whether or not a test taker could perform a specific target task. This finding is identical to that in the researcher's previous pilot study conducted two years ago in the same context.

As introduced in Table 30, the three experts provided more positive feedback on the task-centered rating rubric compared to language-centered rating rubric, as they expect more direct and clear indicators of what test takers can actually do or cannot do in TLU situations. Yet, due to the continuous needs to communicate with U.S. army aviation forces on a regular basis, the experts still prefer to make use of the language-centered rating rubric as an indicator of test takers' colloquial English ability rather than their aviation English in the Korean army context. In terms of revision, MSG Kim suggested that the current task-centered rating sequence and criteria (see Appendix D) need to be modified to be more authentic and natural. Revised task descriptions for task-centered rating rubric are summarized in Figure 9.

<Colloquial Communication>	**<Transition Procedure>**
T-1 Colloquial Communication	T-5-1 Listen to transition request
	T-5-2 Provide traffic info. to a transitioning pilot
<Changed Flight Plan>	**<Arrival Procedure>**
T-2-1 Respond for engine start request	T-6-1 Listen to a pilot's landing request
T-2-2 Respond for flight plan change request	T-6-2 Provide landing info. to a pilot
T-2-3 Listen to changed flight plan	
T-2-4 Take notes of changed flight plan	**<Departure, Arrival, & NOTAM>**
<Terminal Information>	T-7-1 Listen to pilots' departure/landing request
T-3-1 Understand weather info. signs	T-7-2 Provide landing info. to pilots
T-3-2 Listen to weather info. request	T-7-3 Hold aircrafts in complex traffic
T-3-3 Provide weather info. to a pilot	T-7-4 Provide traffic info. to an inbound pilot
T-3-4 Provide departure info. to a pilot	T-7-5 Listen to PIREP
T-3-5 Provide terminal info. to a pilot	T-7-6 Understand NOTAM
T-3-6 Understand terminal info. Signs	T-7-7 Provide NOTAM to a pilot
<Departure Procedure>	
T-4-1 Listen to a pilot's departure request	
T-4-2 Provide terminal info. to a pilot	

Figure 9. Revised task descriptions for task-centered rating criteria

In conclusion, based on the semi-structured phone interview with the three expert army air traffic controllers, the task-centered rating rubric was revised for enhanced appropriateness in providing evidence of test takers' aviation English abilities.

Task Administration Conditions

Research question 2.2 aimed to examine test takers' impressions and opinions about the test administration conditions based on the semi-structured post-test interview questionnaire (see Appendix G). Transcribed interview responses from 16 test takers were analyzed and presented. These responses focused on the authenticity, efficiency, and immersion in Table 31 and satisfaction, comparison with paper-based aviation English test, and suggestions in Table 32.

Test takers' post-test interview responses revealed that military air traffic controllers were positive about the authenticity, efficiency, and immersion of the virtual interactive tasks. Yet, they also addressed issues corresponding to the three categories. The participants' negative comments mainly centered on the interface and design of the simulated air traffic control tower, airbase, and movement of the simulated helicopters in the virtual environment.

Table 31

Authenticity, Efficiency, and Immersion about the Prototype Virtual Interactive Tasks for Aviation English Assessment

	Positive comments	Negative comments
Authenticity	• I agree that the tasks and the simulation features are very authentic. (8) • I felt the simulation tasks very real as I could see the interactive features in the situations.	• Overall atmosphere looks very authentic, but background image and equipment in the tower appear to be more U.S. military style. • It could be more authentic if the tower was equipped with interactive radar monitor. • The movement of aircrafts in the simulation was not like real ones. • In Task 2, we usually approve taxi right away, and then expect the pilot to proceed to the runway by themselves, which was a bit different from what I experienced.
Efficiency	• The task sequence and platform were efficient enough to perform the target tasks. (6) • I was fully able to perform efficiently in the simulated environment. (2)	• Limited visual cues bothered efficient task performance.
Immersion	• I could be fully immersed into the environment. (8) • I became highly motivated and immersed into the environment during the assessment.	• Background engine noise interrupted me. (2) • It provides limited emergence as I had to rely heavily on aural cues rather than visual cues. • Unnatural movement of the avatar bothered immersion. • Anxiety for testing negatively influenced my immersion to the environment.

Regarding test takers' perceptions about immersing characteristics of the simulated ATC tower environment, more than 50% of the test takers responded that they felt fully emerged in the simulated environment and felt highly motivated to perform target tasks. However, the participants also highlighted that limited visual cues in the simulated tower, testing anxiety, and background helicopter noise in the virtual environment could negatively affect test takers' immersion into the authentic target tasks.

First, limited visual cues are indeed one of the greatest limitations of the Second Life simulation environment. As the front view in the ATC tower was primarily pointed towards the middle of runway, other views, especially the right and left side views, were not sophisticatedly designed due to technological limitations. Consequently, when aircrafts take off from and start to leave the front view area, control of the aircrafts becomes difficult and the aircrafts are out of perspective, remaining bigger than they actually should be.

Second, one test taker pointed out unnatural movement of the avatar. It is true that movement of avatars in the simulated environment is not as natural as real human beings. However, the researcher later found that this response was attributed to the earlier Second Life warm-up session conducted with the test takers prior to actual test implementation. During the warm-up session, the test takers were guided to log into the Second Life ATC environment and explore the ATC virtual environment by flying in the air, running, and walking on the runway. This, especially flying in the air, avatar movement might have negatively influenced test takers and led them to be alert that to the fact that this simulation environment represents a surreal space.

Third, there were two test takers who complained about the background helicopter engine noise during the task performance in the simulation environment. This background noise was

intentionally inserted to make the environment seem more authentic, and was accomplished after consulting with expert air traffic controllers. Such complaints seems to be the influence from other standard English listening tests (e.g., TOEIC) practice in which there is no background noise, but rather contains very articulate and loud recordings of native English speakers. This noise issue should be reflected in the warm-up session, so that test takers are not overly distracted by the testing setting.

As for the authenticity of target test tasks and the simulated environment, more than 50% of the test takers agree that the tasks and simulated environment are highly authentic. There are four evaluative comment categories that could be implemented to enhance authenticity: (1) background image and equipment issue, (2) aircraft movement issue, (3) missing radar screen issue, and (4) taxi instruction issue. Among the four issues, (1), (2), and (3) are related to the Second Life simulation, and (4) pertains to the content of the test task.

Test takers pointed out that there are still several areas in the simulated airbase that need to be revised, such as (non-Korean) U.S. military style equipment, unnatural movement of aircrafts, and absence of an interactive radar screen in the ATC tower. In fact, all four issues had been identified prior to actual development of the simulated testing environment with Dr. Sadler, and both he and the researcher spent much time and effort to search for more closely resembling and Korean army-like design items online, but could not find items perfectly fit for the settings. Regarding the interactive radar screen issue, the current simulated tower is, in fact, equipped with a radar screen, but is not interactive with moving aircrafts at the airbase. It would be ideal if the simulated radar could actually display moving aircrafts in the digital world. Yet, this may encompass another level of task design and development in the near future with a sufficient budget and availability of programmers.

In terms of the task content issue, this complaint was anticipated, as explained earlier in discussion of the pilot test task revision (see Table 29). Depending on the size of army airbases, additional departure procedure instructions could be added. Accordingly, test takers who are used to working in such a small airbase might become somewhat confused with the task sequence and task content.

Regarding the issue of efficiency, more than 50% of the test takers agreed that the sequence and platform of the virtual interactive tasks for the aviation English assessment were efficient enough to perform the target tasks. Still, there was still an issue of limited visual cues, a limitation which influenced efficient task performance. Further suggestions are described in Table 32.

Table 32

Overall Satisfaction, Comparison with Paper-based Tests, and Suggestions (*n*=16)

	Comments
Satisfaction	Highly satisfied: 4 Somewhat satisfied: 5 Neutral: 3 (no response: 4)
Compared to paper-based test	• Simulated tasks in Second Life seem to be much better and reliable than paper-based assessment. • Virtual tasks lowered the test taker's burden to imagine the task performance environments, and provide extra. • I would adopt both paper-based tests and virtual task-based tests, as they both have advantages. I would use paper-based for ATC beginners • When I learned and practiced ATC as paper-based, I could feel that I did learn something. However, in a real ATC situation, I found that I could not speak at a loss. I do believe ATC training and testing in a simulated environment can definitely help novice controllers.

Table 32. (continued)

	Compared to paper-based test / training, this simulated environment will help the controllers react to the situation faster and respond to the pilots faster in an authentic way. • As ATC itself is based on interactive communication, I strongly believe that this simulated test will be more reliable and valid than paper-based tests.
Suggestions	• Real-time track display system may need to be incorporated in the tower. • Automated training/testing simulation by adding options, such as selecting nationality of pilots with different English accents, aircraft types, flight mission types, and artificial intelligence can make the tool much better. • The simulated assessment can be more authentic if the simulated task environment can add more airbases with real airbase database including the weather condition, terminal information of individual airbase. • Adding automatic-speech recognition technology generated automated corrective feedback function will help the system much better.

When the test takers were asked about overall satisfaction with the task administration and performance conditions, nine out of 16 responded that they were either highly or somewhat satisfied. These respondents were also asked to compare the virtual interactive tasks for aviation English assessment with conventional paper-based aviation English tests they experienced in the army aviation school. As introduced in Table 32, test takers were highly motivated and satisfied with the simulated task environment, which helped them visualize the target task situation and interactive movement of the aircrafts by reducing their cognitive load, as compared to paper-based tests. Furthermore, respondents also highlighted the advantage of using simulated target task performance settings, which enables test takers to practice and learn air traffic control based on aviation English speaking and listening in a variety of task situations.

To improve the prototype virtual interactive tasks for the aviation English assessment, the participants suggested invaluable ideas. Though the current prototype model of virtual aviation English assessment needs a task operator who controls the sequence and interactive features of the task situations, it is true that there is much room for improvement. One of the most prominent suggestions, one that the researcher has explored over the last three years, is the stand-alone aviation English training and testing system administered with the application of automatic-speech recognition technology and automated corrective feedback. Additionally, the interviewees recommended expanding the simulation environment, including multiple army airbases, so that the test takers could experience ATC in different ATC tower situations using an authentic database of each ATC tower and airbase.

In summary, test takers who experienced the virtual interactive tasks for aviation English assessment thought that the test tasks in the simulated environment were authentic and motivating, in spite of many shortcomings. Depending on the participants' experience in actual air traffic control field work and cutting-edge simulation technology, they seemed to have different standards for, expectations of, and suggestions regarding the current virtual interactive tasks for aviation English assessment. Overall, most test takers expressed that their experience in test administration in the simulated environment through target test tasks was authentic and motivating. Accordingly, it could be expected that their task performance in Second Life can be a good indicator of test takers' real aviation English ability.

Generalization Inference

The generalization inference is based on the warrant that observed scores are estimates of expected scores that test takers would receive across raters. Backing was sought through estimation of inter-rater reliability on test performances.

Systematic Development of Test Specification for Producing Parallel Tasks

Research question 3.1 involved to what extent experts find the test task specification appropriate for producing parallel tasks. To find the answer, three domain experts' perspectives and consensus were investigated through the task-based needs analysis to the pilot test. Although the experts agreed that the current test specification is authentic and ready for future parallel task development, it must be noted that generating a parallel task, especially in a virtual world, requires numerous steps and a great amount of time. Thus, it is necessary to investigate this question fully when technical and financial support can be secured to verify the findings from the current preliminary stage of new test development.

Inter-rater reliability of task-centered rating

Research question 3.1 aimed to investigate inter-rater reliability among each of the two groups of raters who used two different task rating rubrics – a language-centered rating rubric (ICAO, 2012) for aviation English assessment and a task-centered rating rubric developed for this dissertation study and based on the findings from the TLU analysis. The measures used to estimate inter-rater reliability were Cohen's kappa for task-centered rating, because two independent raters rated the same task performance, and Intra-Class Correlation for language-centered rating, because three raters (the population) rated all test takers.

For task-centered rating, two expert air traffic controllers independently assigned a score indicating their judgment of the test taker's task accomplishment level (Excellent=2, Acceptable=1, and Unacceptable=0) to each of the 20 test takers. The statistical measure adopted to calculate inter-rater reliability was Cohen's kappa, which is a measure of the agreement between two raters. The interpretation of the calculated kappa is based on the following range (Lindis & Koch, 1977).

Kappa	Interpretation
< 0	Poor agreement
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Table 33

Inter-rater Reliability of Task-centered Rating using Cohen's kappa

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.588	.033	16.836	.000

The Cohen's kappa is .588 with $p < .05$, which represents a moderate strength of agreement between the two raters who are the expert air traffic controllers. This result can be considered as acceptable agreement, but still less satisfactory compared to substantial or almost perfect agreement. One of the factors affecting the Kappa might be the two raters' disagreement in the unacceptable, acceptable, and excellent levels of task accomplishment. Table 34 shows the number of exact agreements and disagreements for each of the three rating categories.

Table 34

Task-centered Rating Results by Two Expert Controllers

	C-1		C-2		C-3		C-4		C-5		C-6		C-7		C-8		C-9		C-10		C-11		C-12		C-13		C-14		C-15		C-16		C-17		C-18		C-19		C-20	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
Task 1	0	0	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	1	**2**	**1**	0	0	**2**	**1**	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1
Task 2-1	0	0	2	2	1	1	0	0	0	0	**2**	**1**	2	2	0	1	1	1	**2**	**1**	0	0	2	2	0	0	**2**	**1**	0	0	0	0	0	2	0	0	1	0	0	1
Task 2-2	0	0	1	1	1	0	0	0	0	0	0	1	1	1	**1**	**2**	**1**	**2**	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Task 2-3	0	0	0	0	0	0	0	0	1	1	2	2	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Task 2-4	0	0	0	0	0	0	0	0	0	1	2	2	0	0	2	2	**2**	**1**	0	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	1	1
Task 3-1	0	1	1	1	0	0	0	0	2	2	2	2	2	2	1	1	1	1	1	1	2	2	0	0	**1**	**2**	0	1	1	0	2	2	0	0	0	0	0	0	0	0
Task 3-2	0	0	2	2	0	0	0	0	1	1	**1**	**2**	**1**	**2**	1	1	0	1	0	1	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0
Task 3-3	1	1	0	1	1	1	1	1	2	2	1	1	1	1	0	0	0	0	1	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Task 3-4	0	0	0	0	0	0	0	0	**2**	**1**	1	1	**1**	**2**	1	1	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0
Task 3-5	2	2	1	1	0	0	0	0	2	2	2	2	2	2	1	1	1	1	2	2	0	0	1	1	0	0	0	0	2	2	0	0	0	0	0	0	1	1	2	2
Task 3-6	2	2	0	0	0	0	0	0	1	1	2	2	1	1	0	0	1	1	1	0	1	1	1	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0
Task 4-1	2	2	0	1	1	1	0	1	0	1	1	1	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Task 4-2	2	2	1	1	0	0	0	0	0	1	**1**	**2**	0	0	1	1	1	1	0	1	0	0	0	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
Task 5-1	1	1	1	1	0	0	1	1	**1**	**2**	2	2	1	1	1	1	2	2	0	1	0	0	2	2	**2**	**1**	2	2	0	0	2	0	0	0	0	0	2	2	1	1
Task 5-2	2	2	0	0	0	1	0	0	2	2	2	2	1	1	1	1	0	1	2	2	1	1	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
Task 6-1	0	1	0	0	2	2	0	0	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Task 6-2	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	0	0	1	1	0	0	0	1	1	1	0	0	1	0	0	0	0	1
Task 7-1	0	0	0	0	1	1	**1**	**2**	1	1	1	1	1	1	1	2	**1**	**2**	1	1	1	1	0	0	2	2	0	0	1	1	0	0	0	0	**2**	**1**	1	0	0	1
Task 7-2	2	2	0	0	0	0	0	0	1	1	1	1	0	0	1	1	2	2	1	1	0	0	0	0	1	1	**1**	**2**	0	0	0	0	0	0	1	0	0	1	0	0
Task 7-3	**1**	**2**	1	1	1	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0
Task 7-4	2	2	0	0	1	0	0	0	1	1	0	0	1	1	0	0	1	1	1	1	1	1	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	1	0	1
Task 7-5	**1**	**2**	0	0	**1**	**2**	1	1	1	1	**1**	**2**	2	2	2	2	**2**	**1**	1	1	**2**	**1**	2	2	1	0	1	1	0	1	0	0	2	2	0	0	0	0	0	1
Task 7-6	1	1	0	0	1	1	0	0	0	1	1	1	1	1	2	2	0	1	1	1	0	0	0	0	**2**	**1**	0	0	1	0	0	0	0	0	1	0	0	1	0	0
Task 7-7	2	2	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1	1	1	**1**	**2**	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0

Notes. C-#: Controllers (test takers); A: Rater A; B: Rater B; 2: Excellent task completion; 1: Acceptable task completion; 0: Unacceptable task completion

There were 120 rating cases out of 480 in which the two raters disagreed, and 27 of the 120 disagreed rating cases were marked as either acceptable or excellent accomplishment. Despite the rater training and calibration session, a large number of disagreements were identified over the target tasks and test takers. It was expected that the judgment of either excellent or acceptable task accomplishment might be more challenging than the judgment of either successful task completion or failure in the task. However, only 23% of the disagreements

are attributed to the disagreement of either excellent level or acceptable level, and the remaining 77% of the disagreements seem to have originated from the raters' different standards or beliefs about task accomplishment.

The two task-centered raters were asked to provide rationale for their judgment for each task rated. For test taker C-2's task 3-3 (provide weather information to pilots), Rater A rated it as unacceptable and Rater B rated it as an acceptable level of task accomplishment. The following is their rationale:

Rater A: Unacceptable (rationale: the test taker failed to provide weather information to the pilot.)

Rater B: Acceptable (rationale: based on his experience of ATC tower in the past, he provided weather information to the pilot, but needs to improve his English pronunciation.)

According to their rationale, Rater A rated the task mainly focusing on the completion of the given task, while Rater B seemed to consider the background of the test taker in the decision making. The disagreements in the task-centered rating suggest that more extensive and in-depth rater training should be provided before the actual test rating to improve the reliability of ratings.

Inter-rater reliability of language-centered rating

For the language-centered rating, three raters independently assigned holistic scores (ranging from one to six) to 20 test takers' oral responses on the virtual interactive tasks for aviation English assessment. To examine inter-rater reliability, a two-way mixed average measure intra-class correlation coefficient, ICC(3), was calculated.

Table 35

Intra-class Correlation Coefficient of Language-centered Rating

	Intraclass Correlation[a]	**95% Confidence Interval**	**F Test with True Value 0**		**F Test with True Value 0**	
			Value	df1	df2	Sig
Single Measures	.450[b]	[.340, .556]	3.451	119	238	.000
Average Measures	.710[c]	[.607, .790]	3.451	119	238	.000

Note. Two-way mixed effects model where rater effects are random and measures effects are fixed;
a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance;
b. the estimator is the same, whether the interaction effect is present or not;
c. this estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

In Table 35, the result is computed based on an ICC(3) with three raters across 20 test takers. The ICC(3,1) in the first line of single measures of the reliability for a single rater's rating is .45. That means ICC(3,k), which in this case is ICC(2,3) = .71. Therefore, 71% of the variance in the mean of these raters is real due to true score variance. There was a relatively acceptable level of inter-rater reliability with the two-way mixed average measure ICC of .710 with a 95% confidence interval from .607 to .790 ($F(119, 238)=3.451$, $p < .000$), which can be counted as backing for the generalization inference.

Explanation Inference

The explanation inference is based on the warrant that expected scores can be attributed to a construct of aviation English proficiency and integrated abilities for air traffic control. Backing was sought through (a) concurrent correlational studies to investigate whether or not the hypothesized relationships between the VITAEA performance measures and the Pearson's Versant Aviation English Test performance measures are reflected in the data, and (b) discourse analysis of test takers' stimulated recall interviews to investigate whether or not strategies

engaged in during task accomplishment are construct relevant and in accordance with theoretical expectations.

Concurrent Correlational Studies

Research question 4.1 investigated the relationship between test performance on the VITAEA and Pearson's Versant Aviation English Test (VAET). Table 36 summarizes the characteristics of the VIAET and Pearson's Versant Aviation English Test (VAET).

Table 36

Comparison of Characteristics of the VIAET and Pearson's VAET

	Prototype VIAET	Pearson's aviation English test
Test content	Seven scenarios of aviation tasks in the context of Korean army air traffic control tower	Reading, repeat, short answer questions, readbacks, correction and confirmations, and storytelling in the context of general civil aviation; eight sections with a total of 78 test items
Test goal	To provide diagnostic feedback	To rank test takers based on ICAO language proficiency criteria
Test population	Korean army air traffic controllers	Non-native English speaking civil pilots and air traffic controllers
Test delivery	Computer (simulated virtual environment)	Computer or telephone
Test rating	By trained human raters	By automated speech processing technologies
Target construct	Aviation English communication proficiency	Aviation English communication proficiency

Considering the characteristics of the two tests, the researcher adopted the Pearson's VAET for investigating whether or not concurrent correlations between the two measures of the same construct could provide backing for the quality of the VIAET as a measure of Aviation English communication proficiency. Both the new and existing tests share the same target construct of aviation English communication proficiency. The existing test (VAET) was believed to be much better validated with precision with international standards than any other conventional aviation English test.

In order to answer this research question about the concurrent correlations, Pearson correlation coefficients were computed to investigate if there were any systematic relationships between the levels of aviation English abilities as measured by the Pearson's VAET and the VITAEA. Pearson's VAET adopts ICAO's LPR rating which focuses on six constructs (comprehension, fluency, interaction, structure, pronunciation, vocabulary); the language-centered rating rubric of VITAEA was also based on the identical language-centered rating rubric. Additionally, a task-centered rating rubric was also adopted to rate test takers' performance in the VITAEA as a means of measuring their task completion.

As test takers' aviation English performance was rated with the two rating rubrics (i.e., language-centered rubric and task-centered rubric) in the VITAEA, the correlation coefficient was computed for three sets of scores: scores from the VAET and language-centered VITAEA; scores from the VAET and task-centered VITAEA and; scores from the language-centered VITAEA and task-centered VITAEA. Though 20 test takers participated in the VITAEA, only 17 test takers could take the VAET exam due to the discharge and transfer of three air traffic controllers who took the VITAEA.

Regarding the expected relationships between the three sets of scores, it is hypothesized that the test scores of Pearson's VAET and VITAEA (language-centered rating) and it was hypothesized that VITAEA (language-centered rating) and VITAEA (task-centered rating) would be positively correlated. Due to the small sample size, descriptive statistics and graphs are used to display the findings. The descriptive statistics for the three sets of scores are shown in Table 37. Scatterplots of the three test scores are provided in Figure 11. Lastly, the correlation coefficients between the three sets of scores are displayed in Table 38.

Table 37

Descriptive Statistics of Three Test Scores – Pearson's VAET, VITAEA (Language-Centered Ratings) and VITAEA (Task-Centered Ratings)

Score	*N*	*M*	*SD*	*Min*	*Max*
VAET	17	2.66	0.74	1.68	4.23
Language-Centered VITAEA	17	3.68	0.86	2.11	5.11
Task-Centered VITAEA	17	0.6	0.35	0.21	1.69

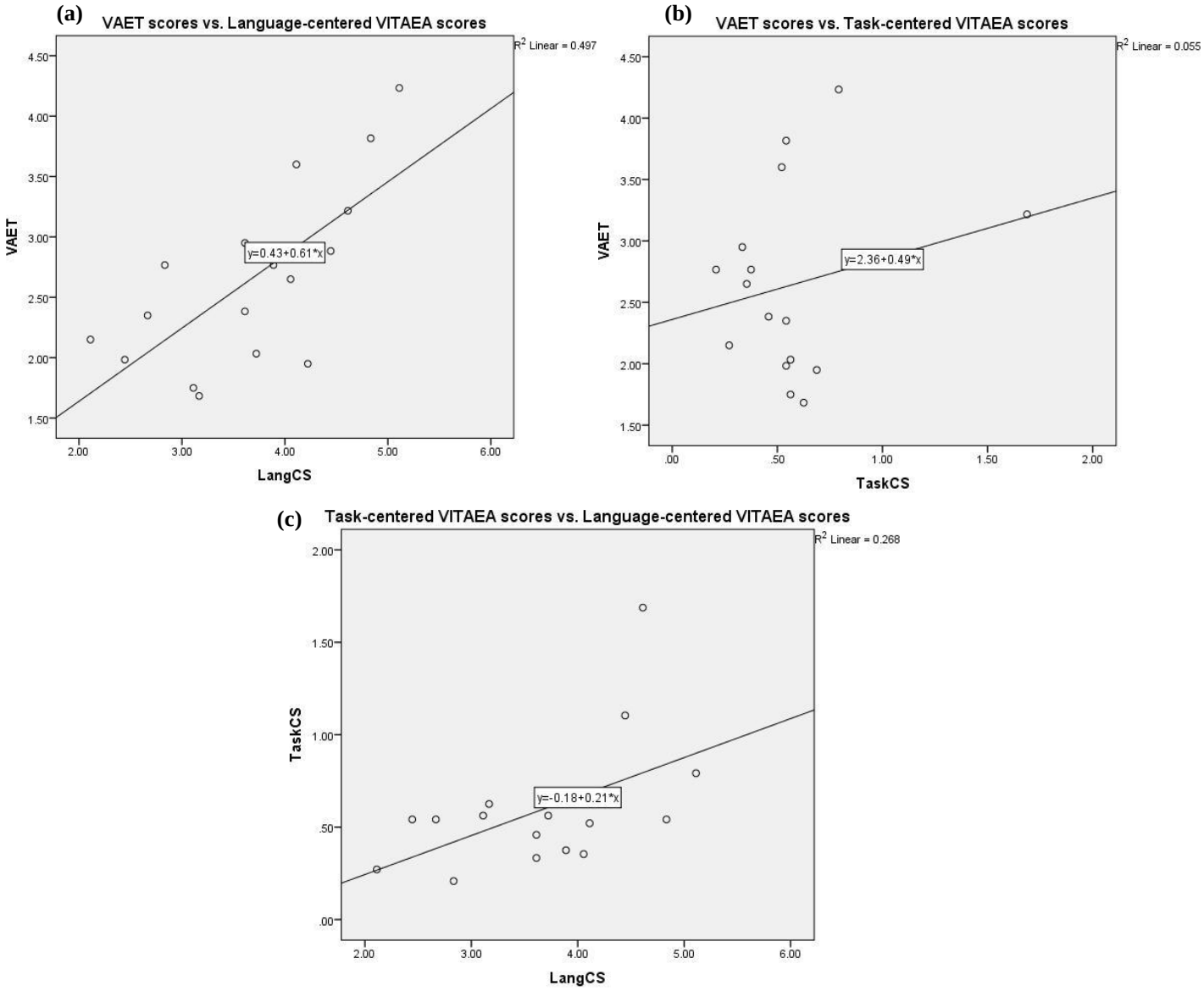

Figure 10. Relationship between test scores of VAET and VITAEA

Notes: VAET: Pearson's Versant Aviation English Test scores; TaskCS: task-centered rating scores of VITAEA; LangCS: language-centered rating scores of VITAEA.

Table 38

Correlation coefficients between Pearson's VAET and VITAEA (Language-Centered Ratings); VAET and VITAEA (Task-Centered Ratings); and VITAEA (Language-Centered Ratings) and VITAEA (Task-Centered Ratings)

		Pearson's VAET	**VITAEA** (language-centered rating)	**VITAEA** (task-centered rating)
Pearson's VAET	Pearson Correlation	1		
	Sig. (2-tailed)			
	N	17		
VITAEA (language-centered rating)	Pearson Correlation	**.710****	1	
	Sig. (2-tailed)	**.002**		
	N	**17**	17	
VITAEA (task-centered rating)	Pearson Correlation	.230	**.520***	1
	Sig. (2-tailed)	.364	**.033**	
	N	17	**17**	17

First, Figure 10(a)'s scatterplot between VAET scores and language-centered scores of VITAEA shows that there is a positive relationship between the two tests. In other words, an air traffic controller who does well in the VAET tasks was also likely to do well in the VITAEA when the task performance was rated with the same language-centered rating rubric. Thus, the backing for the assumption under the explanation inference is somewhat supported.

Second, it appears from Figure 10(b)'s scatterplot between scores from the VAET and task-centered VITAEA that there is no obvious relationship between the test scores of the VAET and the task-centered VITAEA scores. Unlike the positive relationship between the VAET scores and the language-centered VITAEA scores, the negligible relationship may result from

the different construct of interest that was fostered not only with target tasks and task performance environment, but also with the rating rubric used in the assessment. In the case of the VAET, the construct of interest is aviation English communication proficiency, as can be observed through the six constructs of aviation English proficiency. On the other hand, the construct of interest in the virtual interactive tasks for aviation English assessment is the performance of the task itself, which can be accomplished through the use of contextual factors, learner factors (including the six constructs of aviation English language ability), and unidentified constructs. Another interpretation would be that the variance produced by the task-based rubric was considerably smaller, leaving less possibility for covariance among the two tests.

Third, Figure 10(c)'s scatterplot between task-centered rating scores of the VITAEA and language-centered rating score is based on the identical test. It appears from the scatterplot that there is a positive relationship between the VITAEA scores of the task-centered ratings and language-centered ratings. A positive relationship is quite understandable and encouraging, as both the task-centered rating rubric and language-centered rating rubric were designed to measure related, but not identical dimensions of aviation English communication proficiency through observation of task performance. Nonetheless, it is worth noting that those test takers who were able to score relatively well on the language-centered rating were shown to be unsuccessful in the task-centered rating, as revealed in their scoring an unacceptable level of target task accomplishment. To address this discrepancy, the interactionalist construct definition (Chapelle, 1998) needs to be revisited.

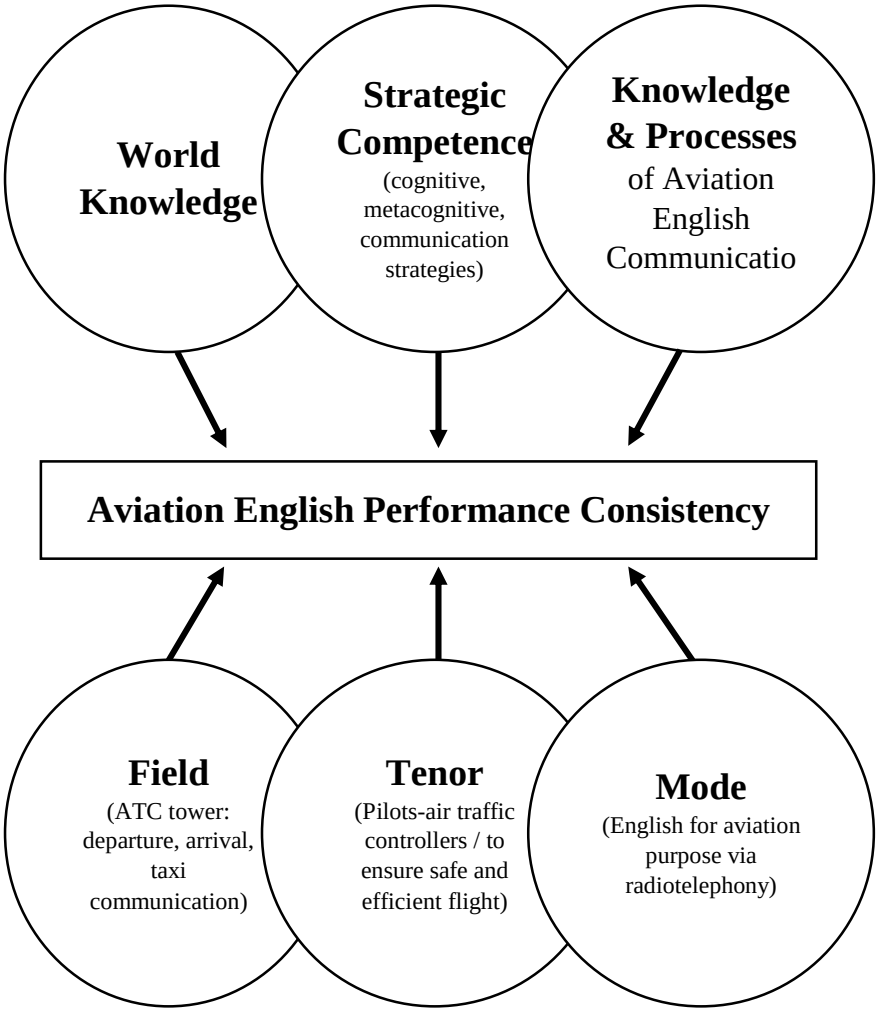

Figure 11. Interactionalist construct Definition of aviation English based on Chapelle (1998, p. 47)

As introduced in Chapter 2, from the perspective of the interactionalist construct definition, aviation English communication abilities in a target language context involve not only the knowledge of aviation English communication and the context of the Korean army ATC situation, but also the strategic competence which directs and assesses its use. In this regard, ICAO's LPR rating criteria, which were adopted for this language-centered rating, primarily focus on the test takers' language knowledge and fundamental processes. While, the task-centered rating rubric was designed to measure task accomplishment. Indeed, such levels of task

accomplishment are directly and indirectly connected with test takers' factors and contextual factors. Although the task-centered rating rubric designed for this dissertation study does not reveal to what specific extent each factor and its sub constructs interplay in the successful or unsuccessful completion of target tasks, findings in response to research question 4.2 still provide some evidence to support that task-centered rating scores do reflect what required factors and their sub constructs seek to accomplish with the given target tasks.

In research question 4.1, as a way to investigate the meaning of the scores from the newly developed virtual interactive tasks for aviation English assessment (VITAEA), concurrent correlations were explored among the two sets of measurements obtained in Pearson's VAET and the VITAEA for the same test takers (n = 17). Despite numerous differences in the two tests, such as format, target task goal, number of task items, task environment, and time duration, the moderate positive correlation between scores from the VAET and language-centered rating of the VITAEA suggests that both tests measure similar constructs of aviation English proficiency to some extent. Nonetheless, considering the differences in the two aviation English tests presented in Table 36, it would be unreasonable to conclude that the prototype VIAET measures exactly the same construct as the VAET. In view of the same language-centered rating rubrics used for both tests, the tasks eliciting examinee performance appear to be responsible for some of the variance in test scores as well.

Strategy Use during the Task Performance

Research question 4.2 aimed to examine (a) what types of aviation English communication strategies were used in the virtual aviation English task performance, and (b) what was the interplay and sequence of these strategies. This is needed as backing for the

assumption that strategies engaged in by test takers are construct relevant and in accordance with theoretical expectations, as expected scores must be attributed to a construct of aviation English proficiency, including appropriate strategy use. To find the answer to this research question, verbal reports from stimulated recalls with the 12 air traffic controllers were transcribed and analyzed. The verbal reports focused on the types of cognitive, metacognitive, and communication strategies used in the virtual interactive tasks for aviation English assessment and their relationship with task performance. Initially, 20 test takers were invited to participate in the stimulated recall data collection; however, due to the quality of audio recording and unexpected absences as a result of emergency duties, 12 test takers' stimulated recall data were analyzed.

Table 39

Identified Strategy Use during Task Performance in VITAEA

Type of Strategy		Sources (Test takers)	Cases	Proportion
Cognitive strategy use	Applying Rules (APR)	12	164	44.09%
	Linking with Prior Knowledge (LPK)	10	60	16.13%
	Inferencing (INF)	10	40	10.75%
	Clarifying (CLAR)	9	20	5.38%
Metacognitive strategy use	Monitoring during performance <MON>	9	35	9.41%
	Self-Evaluation after performance <SE>	9	31	8.33%
	Assessing the situation before performance <ASSIT>	9	22	5.91%

Findings from the stimulated recall data analysis reveal a variety of cognitive and metacognitive strategies that were adopted during the test task performance. *Applying rules* accounts for the largest part (44.09%) followed by *Linking with prior knowledge* (16.13%). Both strategies constitute more than 60% of the entire reported strategy use, which corresponds to the findings from the TLU analysis. As aviation English language consists of standard phraseology,

a simplified, situational, and highly structured version of English used in a very restricted context, test takers may need to utilize these linguistic rules, as well as learned situational rules, in order to perform the target tasks. Additionally, for those who have experienced ATC relatively longer than other test takers, there was a tendency to apply more experience and/or knowledge to the problem solving process.

In addition to test takers' cognitive and metacognitive strategy use, their communication strategies, which were adopted during their test task performance, were coded by two coders (the researcher and an applied linguist) with acceptable intercoder reliability (87.4% agreement). Table 40 displays identified communication strategies adopted by the test takers during the VITAEA task performance.

Table 40

Identified Communication Strategy Use during Task Performance in VITAEA

	Type of Strategy	**Sources** (Test takers who used the strategy)	**Cases**	**Proportion**
Communication strategy use	Direct appeal for help	6	12	18.46 %
	Use of filler	7	10	15.38%
	Omission	7	9	13.85 %
	Asking for repetition	4	8	12.31 %
	Mumbling	5	6	9.23 %
	Indirect appeal for help	3	6	9.23 %
	Self-repair	5	5	7.69 %
	Response confirmation	1	3	4.62 %
	Response rephrase	2	2	3.08 %
	Express non-understanding	1	1	1.54 %
	Message replacement	1	1	1.54 %
	Response expand	1	1	1.54 %
	Response repetition	1	1	1.54 %

Four of the most frequently used communication strategies are *Direct appeal for help* (18.46%) followed by *Use of filler* (15.38%), *Omission* (13.85%), and *Asking for repetition* (12.31%). These four strategies account for about 60% of the total communication strategy use by the test takers. As Douglas (2001) highlighted, communication strategies were indeed employed even when there were no obvious difficulties during the test task performance. Yet, with the given limited number of participants and scarcity of data, there was a tendency for most experienced air traffic controllers to not apply communication strategies as frequently as novice controllers.

Unlike the data set from the 17 test takers used above, a smaller number of participants (*n* = 9) resulted in violating normal distribution of the variables. There were only nine test takers who took both aviation English tests and whose stimulated recall and communication strategy use data collection were completed successfully. As the *N* size is too small to use correlations, the results are presented in Table 41 and Figure 12 as follows.

Table 41

Summary of the Nine Test takers' Strategy Uses and Test Scores

Examinee#	Strategy total	Strategy type	Com-Strategy	TaskCS	LangCS	VAET
3	9	2	7	0.27	2.11	2.15
6	43	6	0	1.69	4.61	3.22
7	36	7	7	0.54	2.44	1.98
8	17	4	6	0.21	2.83	2.77
10	42	8	8	0.79	5.11	4.23
11	42	10	0	1.10	4.44	2.88
13	33	7	0	0.54	4.83	3.82
17	31	8	4	0.56	3.11	1.75
19	26	4	5	0.33	3.61	2.95

Notes. Strategy total: total number of cognitive/metacognitive strategy uses, Strategy type: cognitive/metacognitive strategy types, Com-Strategy: communication strategy use, TaskCS:

task-centered scores of VITAEA, LangCS: language-centered scores of VITAEA, VAET: Versant's Aviation English Test scores

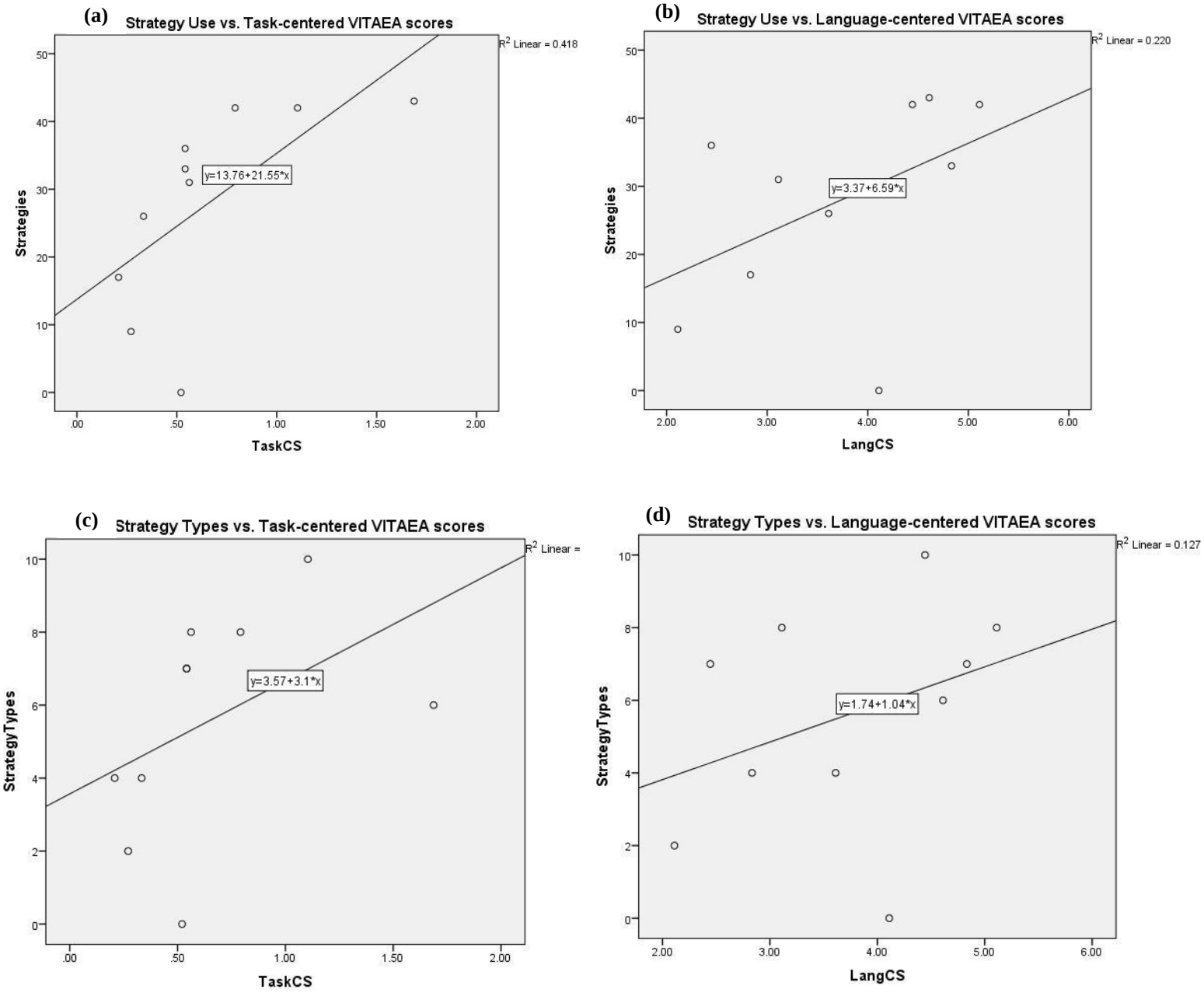

Figure 12. Relationships between VITAEA scores and strategy uses

Despite the limitation of small *N* size and the variables' (test scores and frequency of strategy use/type) unit, overall it appears that a positive relationship exists between test scores and strategy use. According to Figure 12(a) and (b), there appears to be a positive relationship between the total number of strategies use and the two scores on the VITAEA. This may suggest that test takers with higher test scores in the VITAEA tended to use a greater number of cognitive and metacognitive strategies during the test task performance. More significantly, as is apparent in Figure 12(a), a positive relationship between task-centered VITAEA scores and the strategy use may also indicate that those test takers given a high score in the ATC task accomplishment utilized a greater number of cognitive and metacognitive strategies during the task performance. Furthermore, a positive relationship between the total number of strategy types test takers utilized and task-centered scores on the VITAEA also suggests there is a positive relationship between the two variables. Overall, findings from the strategy use study provide some backing for the assumption that strategies engaged by tasks are construct relevant based on the positive relationship between the VITAEA test scores and test takers' strategy use. Moreover, the results are in accordance with the interactionalist construction definition in which strategic competence directs and assesses world knowledge and language knowledge during task performance.

This critical finding of a positive relationship between cognitive/metacognitive strategy use and task-centered scores on the VITAEA provide some empirical evidence that corresponds to Douglas's (2001) emphasis on the test developers' responsibility for "providing sufficient contextual information to enable the test takers to establish the context, to know where they are, and engage an appropriate discourse domain" (p. 76) in the authentic virtual environment by actively utilizing strategic competence. In other respects, the finding could also, to a certain

degree, support the interactionalist construct definition (Chapelle, 1998), which attributes observed performance consistency to the combined influence of learner factors and contextual factors, by providing empirical evidence for the effect of one of the learner factors, strategic competence.

In case of cognitive and metacognitive strategy use, data were extracted from the retrospective recall right after the test to identify test takers' cognitive processes. By contrast, communication strategy use data were gathered from test takers' actual test task performance by conducting a coded analysis of the test takers' oral responses during the test performance to examine what amount and types of communication strategies the test takers adopted while they were trying to accomplish the given tasks. A closer look at the coded communication strategy data reveals that, in many cases, test takers utilized a communication strategy when they were faced with difficulties during the task performance; yet, there were also some cases in which test takers adopted communication strategies when there had been no obvious difficulties. This finding also empirically supports Douglas's (2001) definition of communication strategies, in which he views strategic competence and communication strategies as essential parts of the general language use process, not as simply controlling options for working around the breakdowns in communication or for enhancing language production.

Summary of Results

Table 42 summarizes the results corresponding to the research questions with the data used to answer them.

Table 42

Summary of Results

	Research Question	Backing Supports Assumption	Results for Backing
Domain Description	1.1 Domain analysis - What are the important skills, knowledge, abilities, and processes needed for aviation English communication in the Army Aviation context as identified by expert air traffic controllers and training manual books?	Yes	The researcher identified required knowledge, processes, and skills in the army aviation English context.
	1.2 Domain analysis - What are authentic target tasks that could be representative of the target domain of aviation English communication in the Army Aviation context as identified by the task-based needs analysis conducted with experienced air traffic controllers?	Yes	The researcher identified authentic target aviation English tasks in TLU situations.
	1.3 Systematic process of task design and modeling (expert consensus) - What are experts' perceptions of assessment tasks simulated in Second Life?	Partially	A systematic process of task design and modeling in the virtual environment was engaged by ATC experts through pilot tests.
Evaluation	2.1 Systematic rubric development - What are experts' opinions about the use of construct-centered and task-centered rating rubrics for scoring performance responses?	Yes	Task-centered and language-centered rating rubrics were developed, trialed, and revised based on expert consensus.
	2.2 Trial and revision of task administration conditions - What are test takers' perceptions about the task administration conditions for prompting their use of relevant abilities in a virtual environment?	Partially	Task administration conditions were developed, trialed, and partially revised based on 20 test takers' feedback and suggestions.
	2.3 Rater training and calibration; Test - To what extent can test raters be trained to avoid rating bias in the task-based performance assessment?	Yes	Responses from the five raters indicated the rater training and calibration were successful.
Generalization	3.1 Systematic development of test specification for producing parallel tasks – To what extent did experts find the test task specification appropriate for producing parallel tasks?	Partially	Evidentiary test development under ECD framework including experts' consensus would enable them to produce parallel tasks. Yet, more concrete research would strengthen support.
	3.2 Inter-rater reliability - How high is the inter-rater reliability?	Yes	Cohen's kappa of .59 ($p < .05$) and Intra-Class Correlation of 710 with 95% CI from .607 to .790 indicated good level of reliability.
Explanation	4.1 Concurrent correlational studies - What is the relationship between test performance on the VITAEA and Pearson's Versant Aviation English Test?	Partially	Scatterplots and Pearson correlation coefficient value of 0.71 between scores of VAET and VITAEA indicated positive correlation between the two tests.
	4.2 Investigation of test completion processes using think-aloud protocol and screen capturing - What are test takers' test-taking strategies as identified in a discourse analysis of the think-aloud data?	Yes	Descriptive statistics and scatterplots between VITAEA scores and strategy use suggested that strategies engaged by tasks are construct relevant.

Research question 1.1 examined important skills, knowledge, abilities, and processes needed for aviation English communication in the TLU situations in order to describe target language use situations (target domain) of aviation English. By triangulating the findings from

the task-based needs analysis survey, document analysis of ATC training manual books, and domain experts' (expert military air traffic controllers') consensus, the researcher was able to define required knowledge, processes, and skills in the army air traffic control tower situations. Thus, the researcher concluded that the research question can be answered and the assumption can be supported by the data.

Research question 1.2 investigated authentic target tasks in the TLU situations by conducting an in-depth, task-based needs analysis survey with 81 military air traffic controllers in the target context. Identified target tasks and their sequence corresponded to army ATC manual guidelines; CSM Ryu, the most experienced air traffic controller, agreed with the findings as well. Accordingly, the researcher concluded that the research question can be answered and the assumption can be supported by the data.

Research question 1.3 asked experts' perceptions of assessment tasks simulated in Second Life. After piloting the initial prototype of the VITAEA with two expert air traffic controllers, and in-depth, post-piloting interview was conducted. Detailed feedback on the target tasks and the virtual assessment environments led to intensive revision of the VITAEA. Domain experts agreed that the revised version of the VITAEA showed enhanced authenticity as compared with the previous version. However, due to technological limitations, there were a few areas that could not meet the experts' expectations, such as the use of automatic speech recognition technology and more natural movement of avatars and helicopters in the virtual simulation context. Therefore, the research concluded that the assumption can be partially supported by the data.

Research question 2.1 explored experts' opinions about the use of construct-centered and task-centered rating rubrics for scoring performance responses. In order to identify their

perceptions and areas for improvement, semi-structured interviews with the three domain experts (CSM Ryu, SGM Lee, and MSG Kim) were conducted. Based on their critical advice on the draft rating rubric, the researcher was able to revise the rating rubrics to better meet domain experts' needs and expectations. In conclusion, the researcher can assume that the assumption can be supported by the data.

Research question 2.2 sought to identify test takers' perceptions about the task administration conditions for prompting their use of relevant abilities in a virtual environment. Based on the analysis of semi-structured, post-test interview questionnaire responses from 16 test takers, test takers' perceptions of the simulated task environment, its authenticity and efficiency, test taker satisfaction, a comparison with paper-based aviation English tests, and areas for improvement were investigated. Test takers' perceptions helped the revision process of the VITAEA; however, their recommendations could not be fully realized in the revision process due to technological and logistical limitations. Therefore, the researcher concluded that the assumption can be partially supported by the data.

Research question 2.3 aimed to reveal to what extent raters could be trained to avoid rating bias in the task-based performance assessment. Post-rating interviews and online surveys were conducted with two task-centered raters and three language-centered raters. Their responses were very encouraging and positive, and the statistical measurement of rater reliability also suggested that test rater training and actual rating practice were successful. Therefore, the researcher concluded that the assumption was supported by the data.

Research question 3.1 investigated "To what extent did experts find the test task specification appropriate for producing parallel tasks?" to evaluate whether task and rating specifications are well defined so that parallel tasks can be created. The domain experts thought

that the specification would enable them to produce parallel tasks. The researcher concluded that the assumption is partially supported by the data, though more concrete research would further strengthen the support.

Research question 3.2 investigated inter-rater reliability among the two groups of raters who used two different task rating rubrics: the language-centered rating rubric (ICAO, 2012) for aviation English assessment and the task-centered rating rubric developed for this dissertation study. Cohen's kappa analysis results indicated that there existed substantial agreement between the two task-centered raters. Additionally, an intra-class correlation coefficient was calculated to examine the inter-rater reliability of the three language-centered raters. The result of ICC of 710 also confirmed that there was a somewhat acceptable level of inter-rater reliability. Thus, the researcher concluded that the assumption was supported by the data.

Research question 4.1 examined concurrent validity of the VITAEA computing Pearson correlational coefficients of the test scores from the VITAEA and VAET. The Pearson correlation coefficient value of 0.71 between VAET scores and language-centered scores of VITAEA suggests that there is a positive relationship between the two tests. However, this correlation coefficient finding could only partially provide evidence for concurrent validity. Therefore, the researcher concluded that the assumption is partially supported by the data.

Research question 4.2 explored 12 test takers' verbal reports from the stimulated recalls focusing on the types of cognitive, metacognitive, and communication strategies used in the virtual interactive tasks for aviation English assessment and their relationship with task performance. Findings from the data analysis suggest that there is a positive relationship between the total number of cognitive and metacognitive strategies used and VITAEA scores. Therefore, the researcher concluded that the assumption can be supported by the backing.

CHAPTER 6

CONCLUSION

The present dissertation study demonstrated the development process of aviation English test tasks in a virtual environment and investigated the validity for task-based aviation English performance assessment in the context of Korean Army Aviation. In the current dissertation study, the development and validation of the VITAEA test were based on the four inferences and underlying assumptions in an interpretive argument that developed with reference to argument-based validity (Chapelle et al., 2008; Kane, 2006), evidence-centered design (Mislevy et al., 2002, 2003, 2006), target language use situation analysis for the test development in the language for specific purposes (Douglas, 2001), and task-based language assessment (Long & Norris, 2001; Norris et al., 1998).

Adopting a mixed method with a triangulation design, the current dissertation research included two primary data sources: quantitative data from closed-item questions in questionnaires, test-takers' aviation English test (VAET and VITAEA) performance scores, and coded stimulated recall data; and qualitative data from open-ended questions in questionnaires, interviews, and stimulated recall discourse data. As an initial phase of the study, a task-based needs analysis was conducted with 81 military air traffic controllers in the target context. Findings from the needs analysis on the required knowledge, skills, processes, target tasks, and task procedures in the TLU situations and their specific needs for innovative aviation English testing, which break from the conventional paper-based, vocabulary-focused testing, formed the foundation for authentic task and rating rubric development.

A total of 20 military air traffic controllers completed the prototype virtual interactive tasks for aviation English assessment. Two rater groups, consisting of two raters for task-centered rating

and three raters for language-centered rating, scored each test takers' aviation English performance on the virtual environment tasks. As the prototyping phase of a new aviation English assessment, qualitative data on the contents, interface and procedures of the prototype VITAEA from test takers and TLU experts, raters' opinions on the rater training and their rating processes, and test-takers' stimulated recall discourse were analyzed and presented. As quantitative data, task difficulty was calculated in the test development phase. Additionally, the relationship between test takers' scores from the VITAEA and Pearson's Versant Aviation English Test and another relationship between test takers' strategy use and two different rating criteria in the VITAEA test were analyzed. The validity evidence collected in the various phases of the test development and validation serves as backing for inferences in the interpretive argument as well as invaluable resources for the revision of the prototype VITAEA.

The Validity Argument

As introduced by Chapelle (2008), the overall approach to demonstrate how such evidence serves as backing to support the interpretive argument begins with domain description inference and continues all the way to the implication inference to complete the entire interpretive argument. The following Figure 13, based on Chapelle's (2008, p. 18, 349), offers a concise schematic to overview of the focus of this dissertation study in addition to what the next steps should be to continue investigating the interpretive argument.

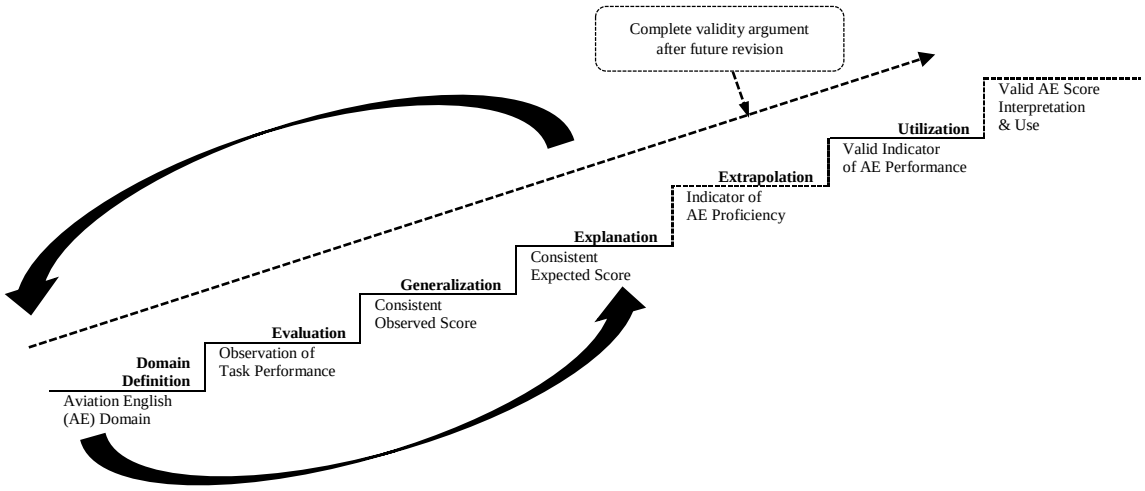

Figure 13. Steps of the VITAEA validity argument based on Chapelle (2008, p. 18, 349)

Figure 13 illustrates a staircase constructed from the steps of the validity argument. Each step represents one inferential bridge that can be crossed with a warrant that is supported by backing. At the bottom of the staircase is the aviation English domain in which aviation English proficiency is needed, and also represents the starting point of the current dissertation study. The cyclical arrows indicate the actual scope of the current study starting from the domain definition to the explanation inference. As the initial prototyping phase of the new aviation English assessment in a virtual environment using authentic target tasks, the researcher expected that all of the findings from the current dissertation study could lead to more evidence-based test design, piloting and revision phases in a cyclical way by supporting the assumptions in each inference and hands-on experience of developing simulated tasks and a rating rubric. As the first prototyping phase of VITAEA, there are a total of four inferences: domain description, evaluation, generalization, and explanation.

Domain Description

As the first inference in the validity argument, the domain description inference links the target language use domain to relevant observable task performance based on the warrant that the observations of performance on the virtual interactive tasks for aviation English assessment are representative of relevant language/topical knowledge of military air traffic control in the context of the Republic of Korea Army Aviation. This warrant is based on the three assumptions that (1) critical aviation English skills, knowledge, and processes needed for aviation English communication in the military ATC context can be identified, (2) assessment tasks that are representative of the English for Specific Purposes (ESP) domain on ATC have been identified, and (3) assessment tasks that require important knowledge, skills, and processes for aviation English communication can be simulated in Second Life.

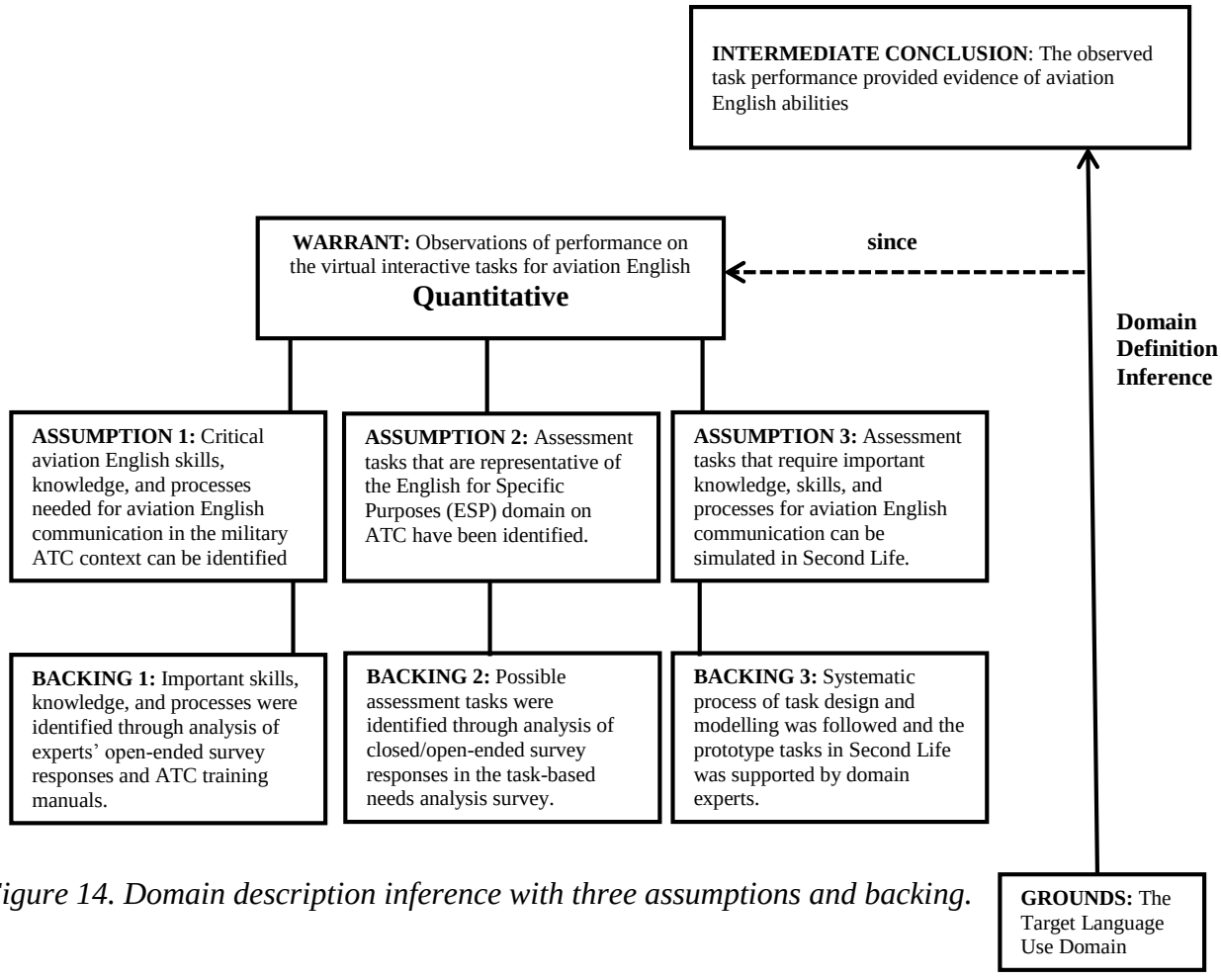

Figure 14. Domain description inference with three assumptions and backing.

Figure 14 presents how the backing supports the domain description inference that links the aviation English use domain to specific performance on the test tasks. As the backing for the first assumption, a domain analysis using open-ended survey responses and a content analysis of the air traffic control manual was conducted; the findings from this first backing informed the description of important skills, knowledge, abilities, and processes needed for aviation English communication. In order to support the second assumption, an in-depth task-based needs analysis survey was conducted with 81 current air traffic controllers in the target context. The findings from the task-based needs analysis informed the development of the target tasks, task situation, and task-centered rating criteria. To endure a systematized process of task design and modeling, qualitative interview data about simulated target tasks in the virtual environment were collected from the domain experts.

Evaluation

The evaluation inference links observations of target task performance to observed scores demonstrating how accurately the observed scores reflect the aviation English communication proficiency (abilities). The evaluation inference is supported by the warrant that observations of task performance on the virtual interactive tasks for aviation English assessment are evaluated to provide observed scores that reflect test takers' aviation English abilities. Three assumptions supporting the evaluation inference are (1) both analytic and holistic rubrics for scoring performance responses are appropriate for providing evidence of targeted language abilities, (2) task administration conditions are appropriate for providing evidence of targeted aviation English ability, and (3) raters can be trained to avoid rater bias.

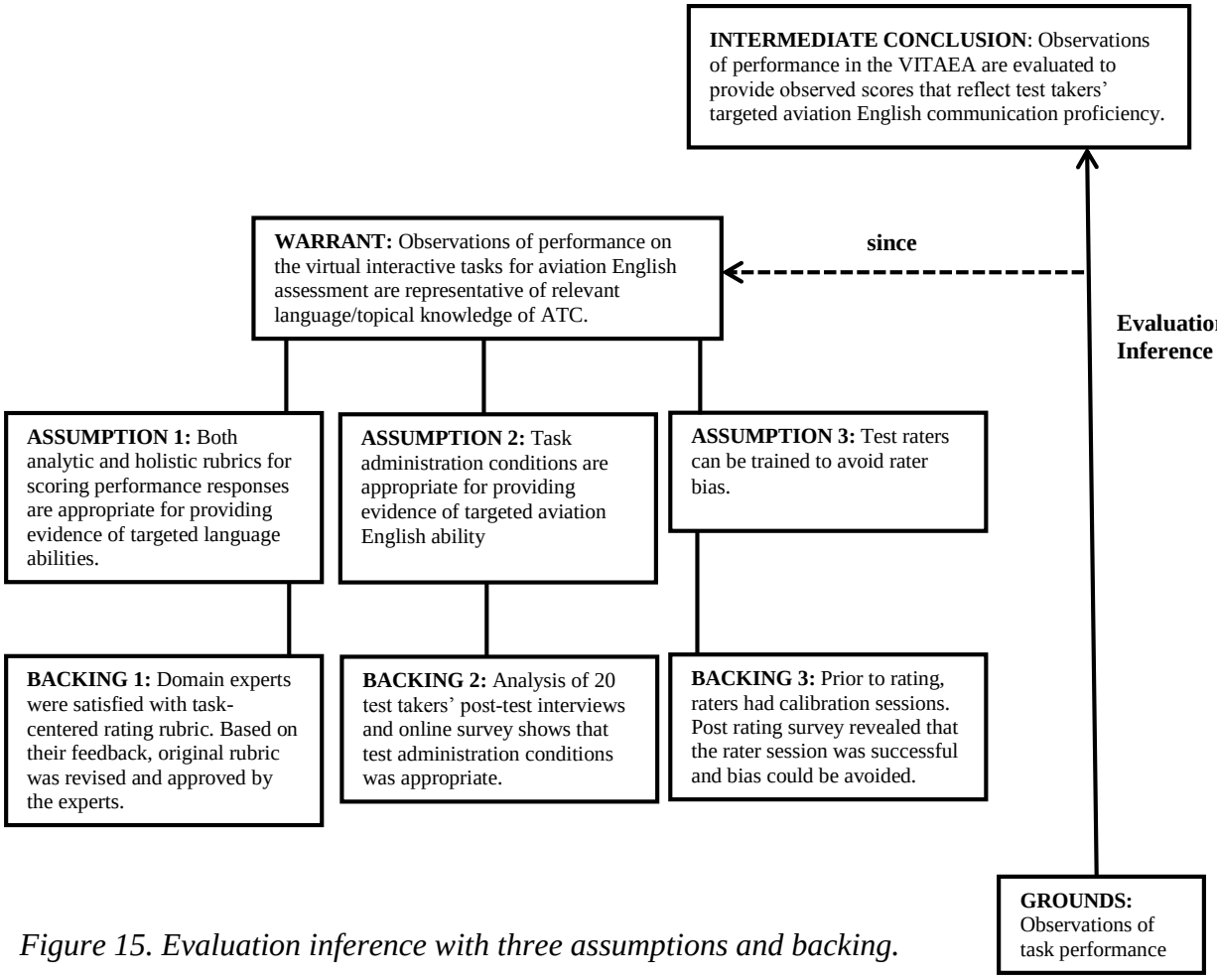

Figure 15. Evaluation inference with three assumptions and backing.

As the very core of the prototype phase of the aviation English test development, these three assumptions are important assumptions to provide evidence of how well test takers' relevant performance in the VITAEA was captured on the tasks and by the raters in the rating rubric. To investigate the three assumptions, three corresponding backing studies were conducted as follows.

For the first assumption about rating rubrics, the backing was found from the qualitative analysis of the two domain experts' interview responses. Both respondents agreed to the proposed task-centered rating rubric and provided additional feedback for minor revisions of the rubric, which were incorporated into the rubric. The backing to support the second assumption

about task administration condition comes from the analysis of 20 test takers' post-test interview and online survey responses. Test takers' responses confirmed that the test administration conditions were appropriate. However, there was also various evaluative feedback recommending follow-up research and task revision with regards to the task scenario, interface of the simulated ATC tower, and more innovative technology use (e.g., automatic speech recognition and automated rating system). The last assumption about rater training is also critical to the evaluation inference. Backing for this third assumption was made with the analysis of a post-rating questionnaire and interview responses from the raters. The raters all confirmed that the rater training and calibration sessions were effective.

Generalization

The generalization inference links observed task performance scores to expected scores based on the warrant that test takers' observed scores are estimates of expected scores on comparable tasks, test forms, administrations, and rating conditions. Two assumptions underlying the warrant includes (a) task and rating specifications are well defined so that parallel tasks can be created; (b) different ratings of different raters are consistent.

The first assumption was supported by the backing that a test specification was systematically developed under the ECD framework for the production of parallel tasks. Although TLU experts thought they could produce parallel task forms, more concrete and in-depth empirical research is needed.

The second assumption was supported by examination of inter-rater reliability. With its initial prototyping phase of a new test development, this dissertation study adopted the assumption that ratings of different raters are consistent. This assumption was supported by the

statistical analysis of inter-rater reliability from the two different rater groups – a language-centered rating group and a task-centered rating group. Cohen's kappa and an intra-class correlation coefficient were adopted to calculate inter-rater reliability of the task-centered rating and language-centered rating respectively. Regarding the results of the backing, substantial agreement was identified in the task-centered rating and a relatively acceptable level of inter-rater reliability was found in the language-centered rating.

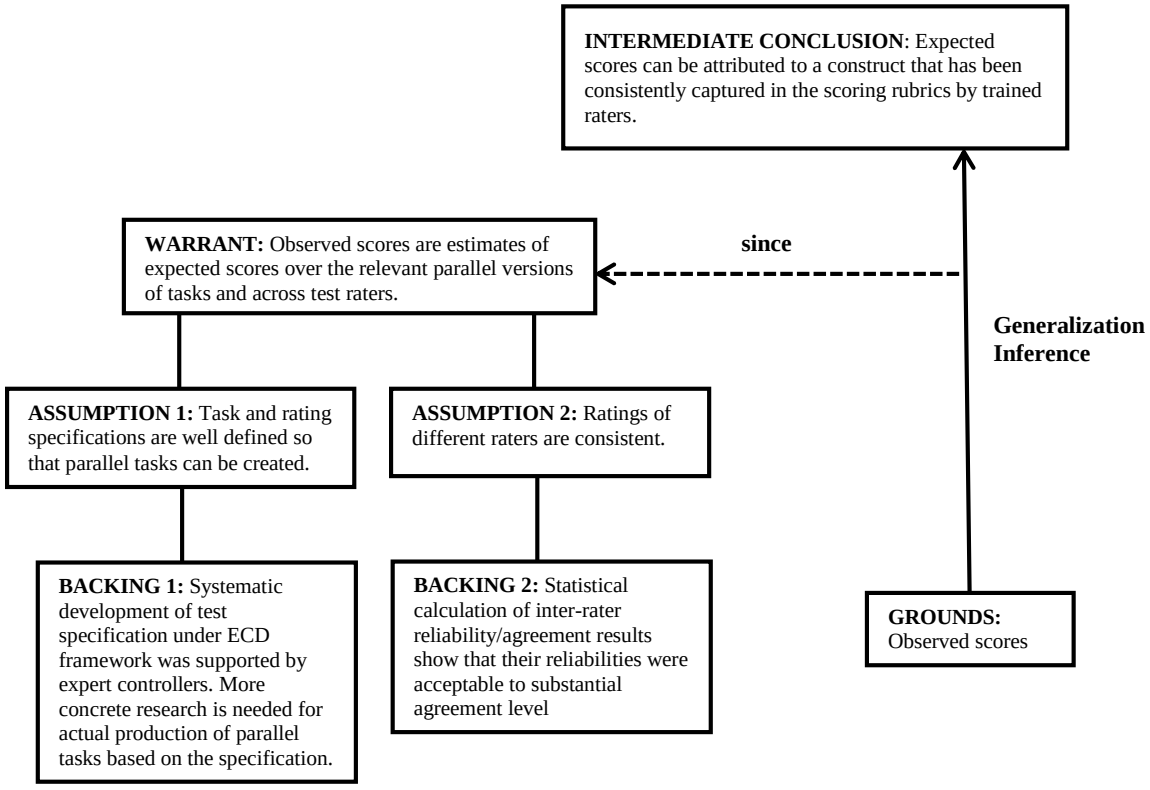

Figure 16. Generalization inference with two assumptions and backing.

Explanation

The explanation inference links expected scores to the construct based on the warrant that expected scores can be attributed to a construct of aviation English proficiency and integrated

abilities for air traffic control. Two assumptions underlying this warrant include (1) performance in the virtual interactive tasks for aviation English assessment relates to performance on other aviation English assessments (e.g., Pearson's Versant Aviation English Test), and (2) strategies engaged by tasks are construct-relevant and in accordance with theoretical expectations.

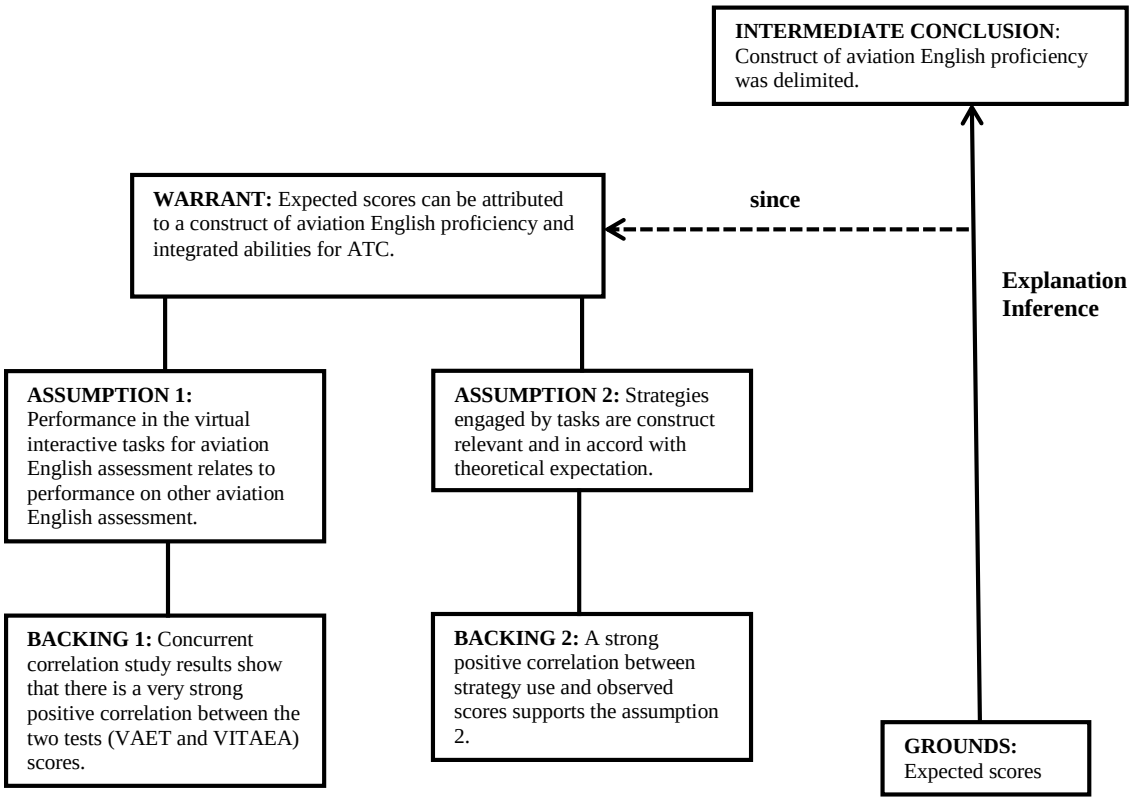

Figure 17. Explanation inference with one assumption and backing.

To support the first assumption, a concurrent correlational study examining the relationship between test performance on the VITAEA and Pearson's Versant Aviation English Test was conducted. The assumption was partially supported by a very strong positive correlation between the two tests scores (i.e., VAET scores and language-centered scores of VITAEA). The second assumption on strategy use was supported by the backing that there was a strong positive correlation between the total number of cognitive and metacognitive strategies used and the two scores on the VITAEA.

Limitations of the Study

The current dissertation study has several limitations. The first limitation concerns the small sample size used for statistical analysis. Although the context was situated in the military airbase and the access to the entire population of military air traffic controllers was limited, only 20 participants took part in the virtual interactive tasks for aviation English assessment. Later in the dissertation study, only nine participants' data could be used for a non-parametric statistical analysis. As the data collection from the task-based needs analysis survey to Pearson's Versant Aviation English Test lasted more than a year and a half, and the military service period of enlisted air traffic controllers is only 21 months, it was very challenging to maintain the same participants throughout the research period. Yet, it would have been beneficial to include more participants in the data collection by conducting various surveys, interviews, and tests in a well-planned timeline.

Second, although this is the first prototyping effort to develop and validate the virtual interactive tasks for aviation English assessments, the current dissertation study could not provide all of the identified backing to support the assumptions under the first four inferences. The study placed more emphasis on the development of the prototype tasks in a virtual environment and evaluative feedback and empirical findings for follow-up revision of the virtual tasks rather than thorough data collection for test validation. What is more, even though the researcher once served in the same ATS Battalion as an air traffic controllers and still has strong communication connections with current military officers, accessing participants in the military in an Asian country and collecting multiple sets of data were not easy tasks for a single researcher to manage.

Third, when it comes to the authenticity of the simulated TLU situations and test tasks, despite the researcher's efforts to create an authentic environment, there is much room for improvement to establish life-like simulations. Test takers still expressed that the air traffic control tower tools and equipment and window view of the air traffic control tower did not meet their authenticity expectations. This is not just an issue of simulation interface, but could be a serious authenticity issue, as all of the equipment and window views are closely connected to air traffic controllers' real aviation English task performance. Controllers do not just rely on listening to a radio or a simple front view of the control tower. This may suggest the fuller context of aviation English proficiency (i.e., supplementing language ability with multiple sources of input). They interact with a binocular telescope, stand, and walk around the tower to secure better views. Additionally, controllers always check real-time track display systems to monitor the inbound and outbound aircrafts and identify unidentified aircrafts. There are numerous critical components that the researcher could not embed in the simulated task environment. Moreover, identified test tasks used for the current dissertation study do not seem to represent the entire TLU situation. The researcher made great efforts to identify, select, and categorize target test tasks from as many as possible authentic tasks. However, there are still numerous, more complex TLU situations and authentic tasks that the current dissertation study failed to account for the level of complexity demanded in a real world ATC TLU situation. The examples may include a situation of bird strike, engine fire, sudden physical issue of a pilot, low fuel of the aircraft, and many more. These TLU situations are not common, but still are very high-stakes scenarios and failure to react appropriately in those TLU situations can lead to fatal injuries or crashes. Bachman (2002) criticizes that in the design of tasks in a task-based approach it is not possible to identify, select, and categorize real-life task types. To some extent, his

skepticism makes sense, as designing target test tasks looks quite challenging. Nevertheless, the endeavor to determine, choose, and classify target tasks and simulate the test task environment in a more authentic way must be worth trying for improving task development.

Implications

Despite the limitations, findings from the current dissertation study have implications for the field of language assessment for language for specific purposes, task-based performance assessment, and computer-assisted language learning (CALL). First, this dissertation study aimed to combine task-based language assessment, evidence-centered design, and argument-based validation for the development and validation of aviation English assessment. For the development of valid interactive aviation tasks, the researcher adopted an argument-based approach (interpretive argument) as a backbone of the framework, synthesized evidence-centered design for valid test development and applied task-based language assessment in the task design phase for task identification, selection, and categorization. Therefore, empirical processes for prototype test development and partial validation based on the theoretical guidance presented in this dissertation study can be seen as one of the first to be constructed utilizing the three theoretical frameworks. Furthermore, this dissertation study can shed light on the steps in applying an argument-based approach for task-based second language assessment. Due to numerous known limitations and issues in task-based language assessment, no known research studies which tried to demonstrate the validation of task-based language assessment were found. The more task-based language assessment is criticized, the more evidential reasoning and argument-based validation must be considered and practiced in the development of task-based second language assessment. It is hoped that this dissertation study will become a useful example

for those who want to develop a valid and reliable task-based language assessment based on validity arguments for their own language testing situations.

Second, despite the great potential for virtual environments, in the field of instructed second language acquisition, the space has been mainly used as an experimental arena for language learning or immersion based on a Vygotskian approach (Schwienhorst, 2002); so, why could these environments not be used for language assessment? The researcher believes that the use of virtual environments could dramatically improve language assessment, especially in language specific purposes assessment, by allowing the observation of test takers' use of situated cognition (cognitive and metacognitive strategies) in addition to collection of their verbal responses. Though the current dissertation study could not distinguish what specific portion of test takers' strategy use accounts for their task accomplishment, this study could identify a very strong positive relationship between test takers' use of cognitive / metacognitive strategies and their task accomplishment scores. This meaningful finding could not have been obtained if the current dissertation study had utilized a paper-based test or a computer-based test with motionless pictures and hyper-text only screen. In this regard, an immersive interface and simulated real TLU situations in virtual environments could provide test takers with more authentic opportunities to perform the target tasks. It is hoped that the researcher's endeavor to utilize virtual environments for LSP assessment can motivate and encourage a so called task-based person, a language-tester, and a CALL person to collaborate in designing and developing more authentic language learning and testing arenas for language learners.

Suggestions for Future Research

The following areas can be further researched in the future. Firstly, further research

should be carried out on the relationship between the two rated scores: language-centered scores and task-centered scores (Brown, 2002). One test taker's performance in this study was analyzed by adopting two different lenses, and the current dissertation study was able to identify that the relationship of the two rated scores were positively correlated. Yet, there could be some meaningful relationship among the six aviation English communication proficiency constructs (i.e., comprehension, fluency, interaction, pronunciation, structure, and vocabulary) and the level of task accomplishment.

More investigation on how coded task difficulty could affect test takers' task accomplishment level will be of great interest. In the present study, difficulty level was coded in identified tasks from the task-based needs analysis by two coders. Although the two-coder reliability was very high and consistent, the researcher failed to identify the relationship between the coded task difficulty and test takers' performance scores. It is expected that test takers would receive lower scores in target tasks of greater difficulty, but in the current dissertation study, evidence for such a hypothesis was not identified. The researcher made another hypothesis in that the concept of test difficulty can be highly subjective. Test takers have different background knowledge and experiences, so tasks with certain difficulty levels might not be perceived by the test takers on the same difficulty level. Empirical research studies are needed to investigate the relationship between the task difficulty level and test takers' performance and also to find alternative ways to code task difficulty depending on the TLU situations or test purpose.

Lastly, the virtual environment for the current dissertation study has much room for improvement especially in technological aspects. Technically, the researcher was already aware of the potential to use automatic speech recognition technology for language assessment, task administration, and evaluative feedback. However, the current dissertation study could not

integrate such technological features into the simulated environment. Future empirical studies on how such innovative technologies can be designed and implemented in virtual environments will be necessary.

REFERENCES

AAS (2012). Air Traffic Control: Army Aviation School.

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing, 27*(1), 51-72.

Amichai-Hamburger, Y., & McKenna, K. (2006). The contact hypothesis reconsidered: Interacting via the Internet. Journal of Computer-Mediated Communication, 11(3), 825–843.

Afflerbach, P. (2000). Verbal reports and protocol analysis. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 163–179). Mahwah, NJ: Erlbaum.

Afflerbach, P., & Cho, B.-Y. (2010). Determining and describing reading strategies: Internet and traditional forms of reading. In W. Schneider & H. Waters (Eds.), *Metacognition, strategy use, and instruction* (pp. 201–225). New York, NY: Guilford.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Ernst Klett Sprachen.

Alderson, J. C. (2006). Final report on a survey of aviation English tests. Lancaster University and the European Organisation for the Safety of Air Navigation (Eurocontrol).

Alderson, J. C. (2009). Air Safety, Language Assessment Policy, and Policy Implementation: The Case of Aviation English. *Annual Review of Applied Linguistics, 29*, 168-187.

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing, 27*(1), 51-72.

Altman, D. G. (1999). *Practical statistics for medical research*. New York, NY: Chapman & Hall/CRC Press.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Annex, ICAO. (2006). to the Convention on International Civil Aviation Aeronautical Telecommunications? Volume II Communication procedures including those with PANS status.

Bachman, L. F. (1990). *Fundamental considerations in language testing*: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453-476.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal, 2*(1), 1-34.

Bachman, L. F., & Palmer, A. (2010). Language assessment in practice: Developing language assessments and justifying their use in the real world: Oxford University Press.

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzan, H. (2005). Making learning fun: Quest Atlantis, a game without guns. Educational Technology Research and Development, 53(1), 86–107.

Barber, C. L. (1962). Some measurable characteristics of modern scientific prose. *Contributions to English syntax and philology*, 21-43.

Barbieri, B. J. (2015). AVIATION ENGLISH: HISTORY AND PEDAGOGY. *Journal of Teaching English for Specific and Academic Purposes, 2*(4), 615-623.

Brown, J. D. (2002). *An investigation of second language task-based performance assessments* (No. 24). University of Hawaii Press.

Brown, J. D. (2004). Performance assessment: Existing literature and directions for research. *Second Language Studies, 22*(2), 91-139.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. Tesol Quarterly, 32(4), 653-675.

Brusso, R. C., Wisher, R. A., Paddock, A., & Hatfield, J. (2014). *Best Practices and Provisional Guidelines for Integrating Mobile, Virtual, and Videogame-Based Training and Assessments*.

Carroll, B. J. (1980). *Testing communicative performance: An interim study*: Janus Book Pub/Alemany Pr.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. E. Enright & J. Jamieson (Eds.) Building a validity argument for the Test of English as a Foreign Language (pp. 319-350). New York, NY: Routledge.

Chapelle, C. (2000). From reading theory to testing practice. Issues in Computer-adaptive Testing of Reading Proficiency. *Studies in Language Testing, 10*, 150-166.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2011). *Building a validity argument for the Test of English as a Foreign Language TM*: Routledge.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple. *Language Testing, 29*(1), 19-27.

Cho, B.-Y. (2014). Competent Adolescent Readers' Use of Internet Reading Strategies: A Think-Aloud Study. Cognition and Instruction, 32(3), 253-289.

Clarke, J. (2006). Making learning meaningful: An exploratory study of multi-user virtual environments in middle school science. Qualifying Paper submitted to the Harvard Graduate School of Education. Cambridge, MA.

Clarke-Midura, J., & Dede, C. (2010). *Assessment, technology, and change. Journal of Research on Technology in Education, 42*(3), 309-328.

Cohen, A. D. (2014). *Strategies in learning and using a second language*: Routledge.

Corbin, J.M., & Strauss, A. C. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. London, UK: Sage.

Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures*. crc Press.

Davies, A. (1999). Dictionary of language testing (Vol. 7). Cambridge University Press.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing, 18*(2), 133-147.

Dede, C. J., Clarke, D., Ketelhut, B., Nelson, B., & Bowman, C., (2005). Fostering motivation, learning, and transfer in multi-user virtual environments. Paper contributed to the American educational research association conference, Montreal, Canada.

Dede, C. (2009). Immersive interfaces for engagement and learning. Science, 323(5910), 66–69.

Dejkunjorn, S. (2005). *Identifying the English Language Needs of Thai Pilots*: Kasetsart University.

Dörnyei, Z., & Scott, M. L. (1997). Communication strategies in a second language: Definitions and taxonomies. *Language learning, 47*(1), 173-210.

Douglas, D. (2000). *Assessing languages for specific purposes*: Cambridge University Press.

Douglas, D. (2001). Language for Specific Purposes assessment criteria: where do they come from? *Language Testing, 18*(2), 171-185.

Douglas, D. (2004). English language testing in the context of Aviation English. *ICAO Journal, 59*(3), 17-18.

Douglas, D. (2014). Nobody seems to speak English here today: Enhancing assessment and training in aviation English. *Iranian Journal of Language Teaching Research, 2*(2), 1-12.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly, 5*(2), 160-167.

Downey, R., Suzuki, M., & Van Moere, A. (2010). High-Stakes English-Language Assessments for Aviation Professionals: Supporting the Use of a Fully Automated Test of Spoken-Language Proficiency. *Professional Communication, IEEE Transactions on, 53*(1), 18-32.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. The Cambridge handbook of expertise and expert performance, 223-242.

Ewer, J. R., Latorre, G., & Derneği, A. N. K. (1969). *A course in basic scientific English* (Vol. 382): Longman London.

Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings. *Online Submission*.

Gerighty, T. (2008). *English for Aviation*. Oxford University Press.

Herbert, A. (1965). *The Structure of technical English. AJ Herbert* (Vol. 208): London]: Longman.

Hines, S. (2010). Evidence-Centered Design: The TOEIC® Speaking and Writing Tests: ETS Research Report No. TC-10-07). Princeton, NJ: Educational Testing Service.

Howard, J. W. (2008). "Tower, Am I Cleared to Land?": Problematic Communication in Aviation Discourse. *Human communication research, 34*(3), 370-391.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes*: Cambridge University Press.

ICAO. (2001). Annex 10. *International Standards and Recommended Practices, Aeronautical Telecommunications: International Civil Aviation Organization., 1*.

ICAO. (2004). Manual on the Implementation of ICAO Language Proficiency Requirements (1st ed.): International Civil Aviation Organization. *International Civil Aviation Organization*.

ICAO. (2007a). Manual on the Implementation of the ICAO Language Proficiency Requirements (2nd ed.): International Civil Aviation Organization. *International Civil Aviation Organization*.

ICAO. (2007b). Manual of radiotelephony: International Civil Aviation Organization.

Jang, H. (2001). *A Study of Improvement on the Army Aviation ATC (Air Traffic Control) English textbook.* (M.A.), Hanyang University, Unpublished master's thesis.

Kalyuga, S. (2007). Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective. Educational Psychology Review, 19(3), 387–399.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. (2006). Validation. In R. Brennen, (Ed.), Educational Measurement (4th Edition), (pp 17-64). Westport, CT: Greenwood Publishing.

Ketelhut, D., Dede, C., Clarke, J., Nelson, B., & Bowman, C. (2008). Studying situated learning in a multi-user virtual environment. In E. Baker, J. Dickieson, W. Wulfeck, & H. O'Neil (Eds.), Assessment of problem solving using simulations (pp. 37–58). Mahweh, NJ: Erlbaum.

Kim, H. (2012). *Exploring the Construct of Aviation Communication: A Critique of the ICAO Language Proficiency Policy* (Unpublished doctoral dissertation). University of Melbourne, Department of Linguistics and Applied Linguistics.

Kolb, D. A. (1984). Experiential learning: Experience as the source of learning and development. New Jersey: Prentice-Hall.

Lamb, G. M. (2006). Real learning in a virtual world. The Christian science monitor, October 5. Retrieved January 3, 2007 from <http://www.csmonitor.com/2006/1005/p13s02-legn.html>.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Leifer, L. (1996). Evaluating product-based learning education. White Paper, Center for Design Research, Stanford University, Palo Alto, CA. Retrieved February 29, 2008 from <http://cdr.stanford.edu/~leifer/publications/Osaka95/Osaka95.ps>.

Lomperis, A. (1996) Nomenclature in ESP. TESOL Matters 6 (2), 10.

Long, M. H., & Norris, J. M. (2000). Task-based teaching and assessment. *Encyclopedia of language teaching*, 597-603.

Long, M. H., & Norris, J. M. (2001). Task-based language teaching and assessment. In M. Byram (Ed.), Encyclopoedia of language teaching. London: Routledge.

Long, M. H. (2005). Methodological issues in learner needs analysis. *Second language needs analysis*, 19-76.

Long, M. H. (2005). Second language needs analysis. Cambridge University Press.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276-282.

McNamara, T. (2012). At last: Assessment and English as a lingua franca. *Plenary talk at 5th International Conference of English as a Lingua Franca*, 24 May, Istanbul.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational Measurement (3rd ed., pp. 60-68). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J., Steinberg, L.S., & Almond, R.G. (2002). Design and analysis in task-based language assessment. Language Testing 19(4), 477-496.

Mislevy, R. J., & Steinberg, L.S. (2003). Focus article: On the structure of educational assessments. Measurement: Interdisciplinary research and perspectives 1(1), 3-62.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 257–305). Westport, CT: American Council on Education/Praeger Publishers.

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military medicine, 178*(10S), 107-114.

Monahan, T., McArdle, G., & Bertolotto, M. (2008). Virtual reality for collaborative e-learning. *Computers and Education, 50*(4), 1339–1353.

Munby, J. (1981). *Communicative syllabus design: A sociolinguistic model for designing the content of purpose-specific language programmes*: Cambridge University Press.

Nguyen, N. T., McFadden, A., Tangen, D. J., & Beutel, D. A. (2013). Video-stimulated recall interviews in qualitative research.

Norris, J., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). Designing second language performance assessments (Technical Report 18). Honolulu: University of Hawaii.

Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing, 19*(4), 337-346.

Norris, J., Hudson, B., Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. Language Testing, 19, 395-418.

Oller, J. W. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die neueren sprachen, 75*(2), 165-174.

Oller, J. W., & Perkins, K. (1980). *Research in language testing*: Newbury House.

Paltridge, B., & Phakiti, A. (Eds.). (2015). *Research Methods in Applied Linguistics: A Practical Resource*. Bloomsbury Publishing.

Pearson, 2008. aah *Versant Aviation English Test: Test description and validation summary*. Palo Alto,CA: Author.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*: National Academies Press.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*: Cambridge University Press.

Rea-Dickins, P. (1987). Testing doctors' written communicative competence: an experimental technique in English for specialist purposes. *Quantitative linguistics, 34*, 185-218.

Richterich, R., & Chancerel, J. (1980). *Identifying the Needs of Adults Leaming a Foreign Language*: Oxford: Pergamon Press.

Robinson, P. C. (1980). *ESP (English for Specific Purposes): the present position*: Pergamon Press.

Sadeghi, K. (2013). Doubts on the validity of correlation as a validation tool in second language testing research: the case of cloze testing. *Language Testing in Asia, 3*(1), 1-17.

Sadler, R. (2011). *Virtual Worlds for Language Learning: From Theory to PRACTICE*. Bern: Peter Lang.

Sadler, R. (2012). Virtual Worlds: An Overview and Pedagogical Examination. *Bellaterra: Journal of Teaching and Learning Language and Literature, 5*(1), 1-22.

Salthouse, T.A. (1991). Expertise as the circumvention of human processing limitations. In K.A. Ericcson & J. Smith (Eds.), Toward a general theory of expertise, (pp. 286-300). Cambridge, England: Cambridge University Press.

Schwienhorst, K. (2002). Why virtual, why environments? Implementing virtual reality concepts in computer-assisted language learning. *Simulation & gaming, 33*(2), 196-209.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practices, 16*(2), 5-24.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428 DOI: 10.1037/0033-2909.86.2.420

Silva, K. T. F. d. (2012). Preparing language teachers to teach in virtual worlds: Analyzing their content, technological, and pedagogical needs.

Skehan, P. (1984). Issues in the testing of English for specific purposes. *Language Testing, 1*(2), 202-220.

Skehan, P. (1996). A framework for the implementation of task based instruction. *Applied Linguistics 17*(1): 38-62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Stevenson, D. K. (1985). Authenticity, Validity and a Tea Party. *Language Testing, 2*(1), 41-47.

Studio, C. (2007). Techsmith. com product website.

Swales, J. M. (1985). *Episodes in ESP*: Pergamon Press.

Taylor, L. (2005). Washback and impact. ELT Journal, 59(2), 154-155.

Toulmin, S. E. (1958), *The Uses of Argument. London*: Cambridge University Press.

Toumlin, S. E. (2003). *The uses of argument* (updated edition). Cambridge, UK: Cambridge University Press.

Turner, J. L. (2014). *Using statistics in small-scale language education research: Focus on non-parametric data*: Routledge.

Van Moere, A. (2010). Automated spoken language testing: Test construction and scoring model development. *Computer-based Assessment (CBA) of Foreign Language Speaking Skills, 84*.

Van Moere, A., Suzuki, M., Downey, R., & Cheng, J. (2011). Implementing ICAO language proficiency requirements in the Versant Aviation English Test. *Australian Review of Applied Linguistics, 32*(3).

Walsh, W. B., & Tyler, L. E. (1989). *Tests and measurements*: Pearson College Div.

Widdowson, H. G. (1983). *Learning purpose and language use*: Oxford University Press Oxford.

Wikipedia. (2014). In Wikipedia, The Free Encyclopedia. Retrieved 04:03, August 10, 2014, from is https://en.wikipedia.org/wiki/Second_Life.

Zokic, M., Boras, D., & Lazic, N.(2012). Computer-aided Aviation English testing on example of RELTA test. In MIPRO, 2012 Proceedings of the 35th International Convention (pp. 1254-1257). IEEE. a.org/wiki/Second_Life

APPENDIX A

VIRTUAL INTERACTIVE AVIATION ENGLISH TASK NEEDS QUESTIONNAIRE

To All Air Traffic Controllers of the Aviation Operations Command:

The purpose of this survey to identify specific target aviation English tasks you have performed and their importance, and your anxiety, attitude, and motivation toward aviation English in the Aviation Operations Command in Korea. This survey is part of our efforts to improve the current aviation English test. Your experiences and opinions are very important to us, and we would greatly appreciate your participation. Thank you very much for your cooperation.

SECTION I: BACKGROUND INFORMATION

1) Current age: _____

2) Gender: _____

3) Current rank: _____
   A. Private Second Class
   B. Private First Class
   C. Corporal
   D. Sergeant
   E. Staff Sergeant
   F. Sergeant First Class
   G. Master Sergeant
   H. Sergeant Major

4) Working position: _____
   A. ATC TWR
   B. Ground Control
   C. FCC
   D. Operation Center
   E. Other (please specify: _____)

5) Years you have served in the military as an air traffic controller: _____

6) Highest level of education you completed: _____
   A. High school
   B. Community college
   C. Four-year college
   D. Graduate school - M.A. program
   E. Graduate school –Ph.D. program

7) Major of highest education: _____

8) Standard English test score: _____
   A. TOEIC
   B. TOEFL
   C. TEPS
   D. IELTS
   E. Other (please specify: _____)

9) Years you have studied general English: _____

10) What were your ways of learning general English? _____

11) Years you have studied aviation English: _____

12) What were your ways of learning aviation English? _____

13) Have you been to any English speaking country before? _____
   A. Yes B. No

14) If your answer to #11 is "Yes", describe which English speaking country(ies) you visited; and the length and purpose of your stay in the country(ies).

15) Describe the aviation English skills (listening, speaking, reading, and writing) and the aviation English knowledge (content) you have been trained about at the Aviation School.

SECTION II: SPECIFIC TARGET AVIATION ENGLISH TASKS

CONTEXT: TOWER CONTROL

1) AVIATION ENGLISH - LISTENING

Please write at least five specific aviation English **LISTENING** tasks you have performed (or trained) for aviation English communication at the ATC tower and indicate the level of its priority extent to which you believe it is important to be assessed in the air traffic control:

1 = Not a priority
2 = Low priority
3 = Slightly Low priority
4 = Slightly High priority
5 = High Priority
6 = Essential

(e.g., listening to a pilot's flight plan change over the radio (5 = High priority))

Specific Aviation English Listening Tasks at the ATC TWR	Priority Scale Low High
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6

2) AVIATION ENLGISH - SPEAKING

Please write at least five specific aviation English **SPEAKING** tasks you have performed (or trained) for aviation English communication at the ATC tower and indicate the level of its priority extent to which you believe it is important to be assessed in the air traffic control:

1 = Not a priority 4 = Slightly High priority

2 = Low priority 5 = High Priority

3 = Slightly Low priority 6 = Essential

(e.g., speaking to a pilot about runway direction for landing (6 = Essential))

Specific Aviation English Speaking Tasks at the ATC TWR	Priority Scale Low High
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6

3) AVIATION ENLGISH - READING

Please write at least five specific aviation English **READING** tasks you have performed (or trained) for aviation English communication at the ATC tower and indicate the level of its priority extent to which you believe it is important to be assessed in the air traffic control:

1 = Not a priority 4 = Slightly High priority

2 = Low priority 5 = High Priority

3 = Slightly Low priority 6 = Essential

(e.g., reading weather condition symbols and abbreviations and understand them (6 = Essential))

Specific Aviation English Reading Tasks at the ATC TWR	Priority Scale Low High
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6

4) AVIATION ENLGISH - WRITING

Please write any specific aviation English **WRITING** tasks you have performed (or trained) for aviation English communication at the ATC tower and indicate the level of its priority extent to which you believe it is important to be assessed in the air traffic control:

1 = Not a priority 4 = Slightly High priority

2 = Low priority 5 = High Priority

3 = Slightly Low priority 6 = Essential

(e.g., writing a flight plan on a flight strip (6 = Essential))

Specific Aviation English Writing Tasks	Priority Scale Low High
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6
	1 – 2 – 3 – 4 – 5 – 6

5) Integrating all four (listening, speaking, reading, and writing) aviation English skills, describe actual sequences of aviation English communication at the TWR. Your specific description of the sequences will help us create authentic aviation English tasks at the TWR.

6) How would you describe the characteristics of **excellent** aviation English communication at the ATC TWR?

7) How would you describe the characteristics of **acceptable** aviation English communication at the ATC TWR?

8) How would you describe the characteristics of **unacceptable** (bad) aviation English communication at the ATC TWR?

SECTION III: ATTITUDE/MOTIVATION TEST BATTERY

6 point Likert scale:
1 - Strongly disagree
2 - Disagree
3 - Somewhat disagree
4 - Somewhat agree
5 - Agree
6 - Strongly agree

1) Attitude Toward a Paper-Based Grammar-Vocabulary focused Aviation English Test

1. A paper-based grammar-vocabulary focused aviation English test is really helpful for aviation English training.

14. I really enjoy taking a paper-based grammar-vocabulary focused aviation English test.

15. A paper-based grammar-vocabulary focused aviation English test is a very important part of aviation English training program.

28. Taking a paper-based grammar-vocabulary focused aviation English test is important because other people will respect me more if I have a good score in the test.

29. I love taking a paper-based grammar-vocabulary focused aviation English test.

2) Attitude Toward a Speaking-Listening focused Aviation English Test in a Virtual Environment

2. A speaking-listening focused aviation English test in a virtual environment is really helpful for aviation English training.

13. I really enjoy taking a speaking-listening focused aviation English test in a virtual environment.

16. A speaking-listening focused aviation English test in a virtual environment is a very important part of aviation English training program.

27. Taking a speaking-listening focused aviation English test in a virtual environment is important because other people will respect me more if I have a good score in the test.

30. I love taking a Paper-Based Grammar-Vocabulary focused Aviation English Test.

3) Anxiety for ATC with Korean Pilots

3. I would get nervous if I had to speak aviation English to Korean pilots.

12. Speaking aviation English to Korean pilots makes me feel worried.

17. It would bother me if I had to speak aviation English with Korean pilots on the radio.

26. I would feel uncomfortable speaking aviation English to Korean pilots.

31. I feel anxious if Korean pilots asks me something in aviation English.

4) Anxiety for ATC with American Pilots

4. I would get nervous if I had to speak aviation English to American pilots.

11. Speaking aviation English to American pilots makes me feel worried.

18. It would bother me if I had to speak aviation English with American pilots on the radio.

25. I would feel uncomfortable speaking aviation English to American pilots.

32. I feel anxious if American pilots asks me something in aviation English.

5) Interest in Aviation English Training

5. I am interested in aviation English training.

10. I have a good feeling toward practicing aviation English.

19. If I had my organization's permission; I would travel to an English speaking country to improve my aviation English.

24. I'd like to live in an English speaking country in the future.

33. If I had some opportunities, I would take private English institute to improve my English.

6) Personality (Extroversion)

6. I get nervous in aviation English practice. (Reverse Coded)

9. During aviation English training, even if I don't understand, I try my best.

20. When I don't understand, I ask the instructor or other colleagues' questions.

23. I like to volunteer answers to questions, regardless of whether I am right or wrong.

34. Making mistakes and being wrong is very embarrassing for me. (Reverse Coded)

7) Organizational Influence

7. Our organization thinks training aviation English is important.

8. Our commanders are interested in aviation English.

21. Good aviation English skills are important for gaining recognition (promotion) in my workplace.

22. Our commanders think that learning aviation English is important.

35. Our commanders expect that I should command a good aviation English.

APPENDIX B

VIRTUAL INTERACTIVE AVIATION ENGLISH TASK NEEDS QUESTIONNAIRE
(Korean version)

항공영어 평가 요구조사 설문지

존경하는 항공관제대대 장병여러분께,

불철주야 국토방위에 헌신하고 계신 관제대대 부사관님과 병사 여러분들의 노고에 진심으로 감사와 존경의 마음을 전합니다. 저도 1999 년 관제대대에서 관제병으로 군복무를 마쳤으며 현재 미국 아이오와 주립대학교 응용언어학과에서 박사과정을 밟고 있는 박문영이라고 합니다. 바쁘신 중에도 불구하고 본 설문조사에 참여해주셔서 대단히 감사드립니다. 본 설문조사는 귀하의 관제영어 세부과업들과 항공영어사용에 대한 귀하의 태도와 동기 분석을 통해 온라인 가상공간에서 항공영어 평가문항을 개발하는데 밑거름이 될 것입니다.

본 설문조사의 예상 소요시간은 30 분정도이며 설문조사는 크게 다음과 같이 세가지 영역으로 나누어져있습니다: 설문참가자 영어학습배경, 항공영어 세부과업, 항공영어에 대한 태도 조사

설문조사에 참가하시는 귀하의 신원은 철저히 지켜질 것이며, 설문참여의 기밀성을 유지하기 위해서 귀하의 이름은 익명으로 처리될 것입니다. 아울러 본 설문의 참여 여부는 귀하의 과업평가에 아무런 영향을 끼치지 않을 것임을 약속 드립니다. 본 설문조사는 자발적인 참여로 이뤄지며 귀하는 언제든 이 설문조사의 참여를 중단하실 수 있으며 그에 따른 어떠한 불이익도 당하지 않음을 밝혀드립니다.

이 설문조사와 관련해서 궁금한 점이 있으면 언제든 아래의 연락처로 문의바랍니다.

- 연구자: 박문영 (이메일: mypark@iastate.edu)

본 설문조사에 소요되는 시간과 구체적 내용에 대해 잘 이해하셨고, 본 설문조사에 자발적으로 참가하시는데 동의하시면 아래 1 번 문항의 빈칸에 귀하의 성함을 기입해주십시오.

SECTION I: 참가자 영어학습 배경조사

1. 설문조사에 참여해주셔서 다시 한번 감사 드립니다. 귀하의 성함과 이메일 주소를 기입해주세요.

(예: 홍길동 / gildong@naver.com) 성함: _____ Email: _____

2. 귀하의 성별: □남성 □여성

3. 귀하의 연령: 만 _____ 세

4. 귀하의 현재 계급: □이병 □일병 □상병 □병장 □하사 □중사 □상사 □원사

5. 귀하의 현재 근무처: □관제타워 (TWR) □FCC □Ground Control (GC) □기타 (_____)

6. 귀하의 관제업무 경력: _____년 _____ 개월

7. 귀하의 최종학력: □고등학교졸업 □전문대휴학 □4 년제대휴학 □전문대졸업 □4 년제대학졸업
□대학원휴학 □대학원졸업 □기타_____

8. 귀하의 최종학력 전공 (대학원 이상인 경우 대학전공 병기): _____

9. 귀하의 최근 공인영어 점수 (복수응답가능):

TOEIC (토익): _____점 TOEFL (토플): _____점 TEPS (텝스): _____점 IELTS: _____점 기타: _____점

10. 귀하의 영어학습 기간: _____년 _____개월 (초중고등학교 기간포함)

11. 귀하의 영어학습 방법: (예: 영어듣기-EBS 교육방송청취하기, 팝송가사받아적기)

영어 말하기	영어 듣기	영어 읽기	영어 쓰기

12. 귀하의 항공영어 학습 기간: _____년 _____개월

13. 귀하의 항공영어 학습 방법:

14. 귀하의 해외 영어연수 경험: □있음 □없음

15. 만약 해외 영어연수 경험이 있으시다면 방문하셨던 국가와 기간:

해외연수 국가: _____ 연수기간: _____년 _____개월

16. 항공관제 후반기 (주특기) 교육 당시 영어의 네가지 (말하기, 듣기, 읽기, 쓰기) 영역에 따른 구체적인 항공영어훈련 (학습) 방법과 내용:

(예: 항공영어듣기–훈련방법: 녹음된 항공영어 교신청취하기–훈련내용: 항공관제 교범의 이착륙 관제)

항공영어	말하기	듣기	읽기	쓰기
훈련(학습)방법				
훈련(학습)내용				

17. 귀하께서 항공관제영어 평가문항을 온라인 가상공간에서 개발하신다면 어떠한 화면구성, 기능, 내용들을 반영해보고 싶으신지 귀하의 의견을 적어주시면 감사하겠습니다.

인터페이스 (화면배치와 구성)	기능	내용

SECTION II: 항공영어 세부과업 조사

► 항공관제 영역: 타워 (TWR) 관제

1. 항공영어 – 듣기

귀하께서 직접 관제타워에서 근무하시면서 혹은 훈련받은 경험을 바탕으로 관제타워에서 담당하셨던 구체적인 항공관제 영어과업 가운데 영어듣기능력과 관련있는 구체적인 과업들을 아래 예시를 참고하셔서 왼쪽 빈칸에 다섯가지 이상 적어주세요. 그리고 나열한 영어듣기 능력과 관련된 항공영어 과업들의 중요도를 오른쪽 척도란에 표시해주세요.

● 중요도 : 1 ---------- 2 ---------- 3 ---------- 4 ---------- 5 ---------- 6 전혀중요하지않음 중요도 낮음 중요도 조금 낮음 중요도 조금 높음 중요도 높음 매우 중요 (필수)
● 항공영어 듣기과업의 예: "비행중 조종사로부터 바뀐 Flight plan 을 무선으로 듣고 받아적기" → 중요도: 1 – 2 – 3 – 4 - 5 - (6) (매우중요(필수))

항공영어 듣기 과업들	중요도 척도 중요도 낮음 ←――――→ 중요도 높음
1.	1 – 2 – 3 – 4 – 5 - 6
2.	1 – 2 – 3 – 4 – 5 - 6
3.	1 – 2 – 3 – 4 – 5 - 6
4.	1 – 2 – 3 – 4 – 5 - 6
5.	1 – 2 – 3 – 4 – 5 - 6

2. 항공영어 – 말하기 (항공관제 영역: 타워 (TWR) 관제)

귀하께서 직접 관제타워에서 근무하시면서 혹은 훈련받은 경험을 바탕으로 관제타워에서 담당하셨던 구체적인 항공관제 영어과업 가운데 영어말하기능력과 관련있는 과업들을 아래 빈칸에 다섯가지 이상 나열해주세요. 그리고 나열한 영어말하기 능력과 관련된 항공영어 과업들의 중요도를 표시해주세요.

● 항공영어 말하기과업의 예: "기지로 착륙하려는 조종사에게 활주로 정보를 무선으로 전달함" → 중요도: 1 – 2 – 3 – 4 - 5 - (6) (중요도 높음)

항공영어 말하기 과업들	중요도 척도 중요도 낮음 ←――――→ 중요도 높음
1.	1 – 2 – 3 – 4 – 5 - 6
2.	1 – 2 – 3 – 4 – 5 - 6
3.	1 – 2 – 3 – 4 – 5 - 6
4.	1 – 2 – 3 – 4 – 5 - 6
5.	1 – 2 – 3 – 4 – 5 - 6

3. 항공영어 – 읽기 (항공관제 영역: 타워 (TWR) 관제)

귀하께서 직접 관제타워에서 근무하시면서 혹은 훈련받은 경험을 바탕으로 관제타워에서 구사하셨던 구체적인 항공관제 영어과업 가운데 영어읽기능력과 관련있는 과업들을 아래 빈칸에 다섯가지 이상 나열해주세요. 그리고 나열한 영어읽기 능력과 관련된 항공영어 과업들의 중요도를 표시해주세요.

● 항공영어 읽기과업의 예: “기상정보와 관련된 기호와 약어들을 읽고 이해하기” → 중요도: 1 – 2 – 3 – 4 - 5 - (6) (중요도 높음)

항공영어 읽기 과업들	중요도 척도 중요도 낮음 ←————→ 중요도 높음
1.	1 – 2 – 3 – 4 – 5 - 6
2.	1 – 2 – 3 – 4 – 5 - 6
3.	1 – 2 – 3 – 4 – 5 - 6
4.	1 – 2 – 3 – 4 – 5 - 6
5.	1 – 2 – 3 – 4 – 5 - 6

4. 항공영어 – 쓰기 (항공관제 영역: 타워 (TWR) 관제)

귀하께서 직접 관제타워에서 근무하시면서 혹은 훈련받은 경험을 바탕으로 관제타워에서 담당하셨던 구체적인 항공관제 영어과업 가운데 영어쓰기능력과 관련있는 구체적인 과업들을 아래 예시를 참고하셔서 왼쪽 빈칸에 다섯가지 이상 적어주세요. 그리고 나열한 영어쓰기 능력과 관련된 항공영어 과업들의 중요도를 오른쪽 척도란에 표시해주세요.

● 항공영어 쓰기과업의 예: “Flight strip 에 운항계획 쓰기” → 중요도: 1 – 2 – 3 – 4 - 5 - (6) (중요도 높음)

항공영어 쓰기 과업들	중요도 척도 중요도 낮음 ←————→ 중요도 높음
1.	1 – 2 – 3 – 4 – 5 - 6
2.	1 – 2 – 3 – 4 – 5 - 6
3.	1 – 2 – 3 – 4 – 5 - 6
4.	1 – 2 – 3 – 4 – 5 - 6
5.	1 – 2 – 3 – 4 – 5 - 6

5. 항공영어의 네가지 영역 (말하기, 듣기, 읽기, 쓰기) 들을 통합한 관제타워에서 실제로 귀하께서 담당하신 (혹은 훈련받으신) 구체적인 항공관제 교신 상황 (시나리오) 들과 그 절차들을 설명해주세요. 예를 들어 기지에서 이륙시, 기지로 착륙시, 공역 통과하는 조종사와의 교신시 세부 절차들 (조종사와 관제사의 요청과 응답)을 설명해주시면 추후 가상공간에서 과업중심의 항공영어문항 시나리오 개발에 큰 도움이 될 것입니다.

5-1. [이륙 관제절차]
5-2. [착륙 관제절차]
5-3. [관제공역 통과]

► 다음은 (6~8) 귀하의 경험을 바탕으로 앞으로 개발될 항공영어 평가문항의 채점기준 (상-중-하)을 파악하고자 합니다. 귀하께서 생각하시는 관제타워에서의 뛰어난 (우수한) (상) / 기준에 겨우 맞는 (중) / 기준에 못미치는 (하) 항공관제영어의 특징을 아래 빈칸에 구체적으로 설명해주세요. 타워관제 상황에서의 항공영어의 평가의 기준 (척도)들을 아래의 항목들을 참고해주시거나 혹은 그외에도 귀하께서 생각하시는 중요한 기준들을 아래 빈칸에 적어주시면 대단히 감사하겠습니다.

[관제타워에서의 항공영어과업의 평가기준들의 예]
- 영어발음의 정확성, - 영어발음의 유창성, - 영어 청취능력, - 관제영어 교신시 반응 속도
- 조종사 요청에 정확하게 정보 제공, - 전체적인 운항흐름 파악하기, - 항공영어 문장의 정확성
- 항공영어 문장의 유창성, - 성공적인 착륙관제, - 성공적인 이륙관제, - 성공적인 기상정보제공

6. 귀하가 생각하는 관제타워에서의 뛰어난 (우수한) 항공관제영어의 특징을 설명해주세요. 다시말해 오랜 경험과 뛰어난 자질을 갖추신 항공관제사들이 구사하는 항공관제영어의 자질(특징)들을 최대한 상세하게 구체적으로 묘사해주시면 감사하겠습니다.
(예: 조종사의 요청에 대한 응답 속도가 빠르고 요청한 정보를 정확하게 이해하고 필요한 정보를 정확하게 제공한다.)

7. 귀하가 생각하는 관제타워에서의 기준에 겨우 맞는 항공관제영어의 특징을 설명해주세요. 다시말해 항공관제 전문가들이 아닌 항공관제 경험이 적은 초보 관제사들이 구사하는 항공관제영어의 자질들을 최대한 상세하게 구체적으로 묘사해주시면 감사하겠습니다.

8. 귀하가 생각하는 관제타워에서의 기준에 못미치는 (틀린) 항공관제영어의 특징을 설명해주세요. 다시말해 항공관제 경험이 거의 없는 초보 관제사들이 제대로 구사하지 못하는 서투른 (불충분한) 항공관제영어의 자질들을 최대한 상세하게 구체적으로 묘사해주시면 감사하겠습니다.

► 항공관제 영역: Flight Coordination Center (FCC) 관제

9. 항공영어 – 듣기

귀하께서 직접 FCC 에서 근무하시면서 혹은 훈련받은 경험을 바탕으로 관제타워에서 구사하셨던 구체적인 항공관제 영어과업 가운데 영어듣기능력과 관련있는 구체적인 과업들을 왼쪽 빈칸에 다섯가지 이상 적어주세요. 그리고 나열한 영어듣기 능력과 관련된 항공영어 과업들의 중요도를 오른쪽 척도란에 표시해주세요.

FCC 에서의 항공영어 듣기 과업들	중요도 척도 중요도 낮음 ←————→ 중요도 높음
1.	1 – 2 – 3 – 4 – 5 -6
2.	1 – 2 – 3 – 4 – 5 -6
3.	1 – 2 – 3 – 4 – 5 -6
4.	1 – 2 – 3 – 4 – 5 -6
5.	1 – 2 – 3 – 4 – 5 -6

10. 항공영어 – 말하기 (항공관제 영역: FCC 관제)

귀하께서 직접 FCC 에서 근무하시면서 혹은 훈련받은 경험을 바탕으로 FCC 에서 구사하셨던 구체적인 항공관제 영어과업 가운데 영어말하기능력과 관련있는 구체적인 과업들을 왼쪽 빈칸에 다섯가지 이상 적어주세요. 그리고 나열한 영어말하기 능력과 관련된 항공영어 과업들의 중요도를 오른쪽 척도란에 표시해주세요.

FCC 에서의 항공영어 말하기 과업들	중요도 척도 중요도 낮음 ←————→ 중요도 높음
1.	1 – 2 – 3 – 4 – 5 -6
2.	1 – 2 – 3 – 4 – 5 -6
3.	1 – 2 – 3 – 4 – 5 -6
4.	1 – 2 – 3 – 4 – 5 -6
5.	1 – 2 – 3 – 4 – 5 -6

11. 항공영어 – 읽기 (항공관제 영역: FCC 관제)

귀하께서 직접 FCC 에서 근무하시면서 혹은 훈련받은 경험을 바탕으로 FCC 에서 구사하셨던 구체적인 항공관제 영어과업 가운데 영어읽기능력과 관련있는 구체적인 과업들을왼쪽 빈칸에 다섯가지 이상 적어주세요. 그리고 나열한 영어읽기 능력과 관련된 항공영어 과업들의 중요도를 오른쪽 척도란에 표시해주세요.

FCC 에서의 항공영어 읽기 과업들	중요도 척도 중요도 낮음 ←————→ 중요도 높음
1.	1 – 2 – 3 – 4 – 5 -6
2.	1 – 2 – 3 – 4 – 5 -6
3.	1 – 2 – 3 – 4 – 5 -6
4.	1 – 2 – 3 – 4 – 5 -6
5.	1 – 2 – 3 – 4 – 5 -6

12. 항공영어 – 쓰기 (항공관제 영역: FCC 관제)

귀하께서 직접 FCC 에서 근무하시면서 혹은 훈련받은 경험을 바탕으로 FCC 에서 구사하셨던 구체적인 항공관제 영어과업 가운데 영어쓰기능력과 관련있는 구체적인 과업들을 왼쪽 빈칸에 다섯가지 이상 적어주세요. 그리고 나열한 영어쓰기 능력과 관련된 항공영어 과업들의 중요도를 오른쪽 척도란에 표시해주세요.

FCC 에서의 항공영어 쓰기 과업들	중요도 척도 중요도 낮음 ←————→ 중요도 높음
1.	1 – 2 – 3 – 4 – 5 -6
2.	1 – 2 – 3 – 4 – 5 -6
3.	1 – 2 – 3 – 4 – 5 -6
4.	1 – 2 – 3 – 4 – 5 -6
5.	1 – 2 – 3 – 4 – 5 -6

13. 다음은 (13-1, 13-2, 13-3) 항공영어의 네가지 영역 (말하기, 듣기, 읽기, 쓰기) 들을 통합한 질문으로 FCC 에서 실제로 귀하께서 담당하신 (혹은 훈련받으신) 구체적인 항공관제 교신 상황 (시나리오) 들과 그 절차들을 기준으로 응답해주시기 바랍니다. FCC 공역을 통과하는 조종사와의 교신시 세부 절차들 (조종사와 관제사의

요청과 응답)을 설명해주시면 추후 가상공간에서 과업중심의 FCC 항공영어문항 시나리오 개발에 큰 도움이 될 것입니다.

13-1. FCC 에서의 관제상황 1:
13-2. FCC 에서의 관제상황 2:
13-3. FCC 에서의 관제상황 3:

14. 귀하가 생각하는 FCC 에서의 **뛰어난 (우수한)** 항공관제영어의 특징을 설명해주세요. 다시말해 항공관제 전문가들이 구사하는 항공관제영어의 자질들을 구체적으로 묘사해주시면 감사하겠습니다.

15. 귀하가 생각하는 FCC 에서의 **기준에 겨우 맞는** 항공관제영어의 특징을 설명해주세요. 다시말해 항공관제 전문가들이 아닌 항공관제 경험이 적은 초보 관제사들이 구사하는 항공관제영어의 자질들을 구체적으로 묘사해주시면 감사하겠습니다.

16. 귀하가 생각하는 FCC 에서의 **기준에 못미치는 (틀린)** 항공관제영어의 특징을 설명해주세요. 다시말해 항공관제 경험이 거의 없는 초보 관제사들이 제대로 구사하지 못하는 서투른 (불충분한) 항공관제영어의 자질들을 구체적으로 묘사해주시면 감사하겠습니다.

SECTION III: 항공영어에 대한 태도

아래의 문항들은 항공영어구사와 항공영어평가에 대한 귀하의 태도와 동기를 파악하고자 합니다. 정답과 오답이 따로 없는 문항들이며 아래의 척도를 참고해서 가능한한 정확하게 응답해주시면 감사하겠습니다.

● 척도 : 1 ----- 2 ----- 3 ----- 4 ----- 5 ----- 6

전혀 동의하지않음 동의하지않음 약간 동의하지않음 약간 동의함 동의함 매우 동의함

1. 문법과 단어위주의 서면 (paper-based) 항공영어평가는 도움이 된다.	1	2	3	4	5	6
2. 온라인 가상공간에서 말하기 듣기 중심의 항공영어평가는 도움이 될것이다.	1	2	3	4	5	6
3. 나는 한국인 조종사들과 관제교신을 하는것이 부담된다.	1	2	3	4	5	6
4. 나는 미국인 조종사들과 관제교신을 하는것이 부담된다.	1	2	3	4	5	6
5. 나는 항공영어훈련 (학습) 에 관심이 있다.	1	2	3	4	5	6
6. 나는 항공영어훈련하는 것이 부담된다.	1	2	3	4	5	6
7. 관제영어숙달은 항작사 (관제대대) 에서 중요하다.	1	2	3	4	5	6
8. 내 지휘관은 항공영어숙달에 관심이 많다.	1	2	3	4	5	6
9. 비록 잘 이해못하더라도 나는 항공영어숙달을 위해 노력하는 편이다.	1	2	3	4	5	6
10. 나는 항공영어학습에 대해 좋게 생각한다.	1	2	3	4	5	6
11. 미국인 조종사들에게 무선 교신하는 것은 가끔 걱정스럽다.	1	2	3	4	5	6
12. 한국인 조종사에게 무선 교신하는 것은 가끔 걱정스럽다.	1	2	3	4	5	6
13. 나는 온라인가상공간에서 항공영어훈련을 받아보고 싶다.	1	2	3	4	5	6
14. 나는 문법-어휘 위주의 항공영어훈련을 받아보고 싶다.	1	2	3	4	5	6
15. 문법-어휘 위주의 항공영어평가는 항공영어학습에 매우 중요한 부분이다.	1	2	3	4	5	6
16. 온라인 가상공간에서의 말하기-듣기 위주의 항공영어평가는 항공영어학습에 중요한 부분이다.	1	2	3	4	5	6
17. 나는 한국인 조종사들과 무선교신 하는 것이 불편하다.	1	2	3	4	5	6
18. 나는 미국인 조종사들과 무선교신 하는 것이 불편하다.	1	2	3	4	5	6
19. 소속기관에서 허락한다면 나는 영어실력향상을 위해 영어권국가로 연수를 가보고 싶다.	1	2	3	4	5	6
20. 나는 교육중 이해가 잘 되지 않는다면 교관님 혹은 동료들에게 질문을 하는 편이다.	1	2	3	4	5	6
21. 훌륭한 항공영어구사능력은 내 근무지에서 인정받고 승진하는데 중요한 편이다.	1	2	3	4	5	6
22. 내 지휘관은 항공영어학습이 중요하다고 생각한다.	1	2	3	4	5	6
23. 정답의 옳고 그름을 떠나서 나는 자발적으로 질문에 답하는것을 좋아한다.	1	2	3	4	5	6
24. 나는 미래에 영어권 국가에 방문하거나 체류하기를 희망한다.	1	2	3	4	5	6
25. 나는 미국인 조종사들과 영어로 교신하는 것이 불편하다.	1	2	3	4	5	6
26. 나는 한국인 조종사들과 영어로 교신 하는 것이 불편하다.	1	2	3	4	5	6
27. 말하기-듣기에 초점을 맞춘 가상공간에서의 항공영어평가는 중요하다. 왜냐하면 그 평가에서 좋은 점수를 얻게 되면 다른 사람들이 나를 존중해 주기 때문이다.	1	2	3	4	5	6
28. 영어단어-문법에 초점을 맞춘 서면 (paper-based) 항공영어평가는 중요하다. 왜냐하면 그 평가에서 좋은 점수를 얻게 되면 다른 사람들도 존중해 주기 때문이다.	1	2	3	4	5	6
29. 나는 문법-어휘에 초점을 맞춘 서면 (paper-based) 항공영어평가를 치르기를 희망한다.	1	2	3	4	5	6
30. 나는 말하기-듣기에 초점을 맞춘 온라인 가상공간에서 항공영어평가를 치르기를 희망한다.	1	2	3	4	5	6
31. 나는 한국인 조종사가 무선으로 요청을 할 때마다 염려한다.	1	2	3	4	5	6
32. 나는 미국인 조종사가 무선으로 요청을 할 때마다 염려한다.	1	2	3	4	5	6
33. 나는 기회가 된다면 영어실력향상을 위해 영어과외나 학원수업을 듣고 싶다.	1	2	3	4	5	6
34. 실수하거나 정답을 틀리는 것은 나를 매우 부끄럽게 만든다.	1	2	3	4	5	6
35. 내 지휘관은 내가 훌륭한 항공영어를 구사하기를 희망할 것이다.	1	2	3	4	5	6

설문조사는 여기까지입니다. 다시한번 설문조사에 참여해주셔서 대단히 감사드립니다.

APPENDIX C

ICAO'S LANGUAGE PROFICIENCY REQUIREMENTS RATING CRITERIA

(LANGUAGE-CENTERED RATING CRITERIA)

Level	**Pronunciation Structure** Assumes a dialect and/or accent intelligible to the aeronautical community.	**Structure** Relevant grammatical structures and sentence patterns are determined by language functions appropriate to the task.	**Vocabulary**	**Fluency**	**Comprehension**	**Interactions**
Expert 6	Pronunciation, stress, rhythm, and intonation, though possibly influenced by the first language or regional variation, almost never interfere with ease of understanding.	Both basic and complex grammatical structures and sentence patterns are consistently well controlled.	Vocabulary range and accuracy are sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics. Vocabulary is idiomatic, nuanced, and sensitive to register.	Able to speak at length with a natural, effortless flow. Varies speech flow for stylistic effect, e.g. to emphasize a point. Uses appropriate discourse markers and Connector spontaneously.	Comprehension is consistently accurate in nearly all contexts and includes comprehension of linguistic and cultural subtleties.	Interacts with ease in nearly all situations. Is sensitive to verbal and non-verbal cues, and responds to them appropriately.
Extended 5	Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, rarely interfere with ease of understanding.	Basic grammatical structures and sentence patterns are consistently well controlled. Complex structures are attempted but with errors which sometimes interfere with meaning.	Vocabulary range and accuracy are sufficient to communicate effectively on common, concrete, and work-related topics. Paraphrases consistently and successfully. Vocabulary is sometimes idiomatic.	Able to speak at length with relative ease on familiar topics, but may not vary speech flow as a stylistic device. Can make use of appropriate discourse markers or connectors.	Comprehension is accurate on common, concrete, and work-related topics and mostly accurate when the speaker is confronted with a linguistic or situational complication or an unexpected turn of events. Is able to comprehend a range of speech varieties (dialect and/or accent) or registers.	Responses are immediate, appropriate, and informative. Manages the speaker/listener relationship effectively.
Operational 4	Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation but only sometimes interfere with ease of understanding.	Basic grammatical structures and sentence patterns are used creatively and are usually well controlled. Errors may occur, particularly in unusual or unexpected circumstances, but rarely interfere with meaning.	Vocabulary range and accuracy are usually sufficient to communicate effectively on common, concrete, and work related topics. Can often paraphrase successfully when lacking vocabulary in unusual or unexpected circumstances.	Produces stretches of language at an appropriate tempo. There may be occasional loss of fluency on transition from rehearsed or formulaic speech to spontaneous interaction, but this does not prevent effective communication. Can make limited use of discourse markers or connectors. Fillers are not distracting.	Comprehension is mostly accurate on common, concrete, and work-related topics when the accent or variety used is sufficiently intelligible for an international community of users. When the speaker is confronted with a linguistic or situational complication or an unexpected turn of events, comprehension may be slower or require clarification strategies.	Responses are usually immediate, appropriate, and informative. Initiates and maintains exchanges even when dealing with an unexpected turn of events. Deals adequately with Apparent misunderstandings by checking, confirming, or clarifying.
Pre-Operational 3	Pronunciation, stress, rhythm, and intonation are influenced by the first language or regional variation and frequently interfere with ease of understanding.	Basic grammatical structures and sentence patterns associated with predictable situations are not always well controlled. Errors frequently interfere with meaning.	Vocabulary range and accuracy are often sufficient to communicate on common, concrete, or work-related topics but range is limited and the word choice often inappropriate. Is often unable to paraphrase successfully when lacking vocabulary.	Produces stretches of language, but phrasing and pausing are often inappropriate. Hesitations or slowness in language processing may prevent effective communication. Fillers are sometimes distracting.	Comprehension is often accurate on common, concrete, and work related topics when the accent or variety used is sufficiently intelligible for an international community of users. May fail to understand a linguistic or situational turn of events.	Responses are sometimes immediate, appropriate, and informative. Can initiate and maintain exchanges with reasonable ease on familiar topics and in predictable situations. Generally inadequate when dealing with an unexpected turn of events.
Elementary 2	Pronunciation, stress, rhythm, and intonation are heavily influenced by the first language or regional variation and usually interfere with ease of understanding.	Shows only limited control of a few simple memorized grammatical structures and sentence patterns.	Limited vocabulary range consisting only of isolated words and memorized phrases.	Can produce very short, isolated, memorized utterances with frequent pausing and a distracting use of fillers to search for expressions and to articulate less familiar words.	Comprehension is limited to isolated, memorized phrases when they are carefully and slowly articulated.	Response time is slow, and often inappropriate. Interaction is limited to simple routine exchanges.
Pre-Elementary 1	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.	Performs at a level below the Elementary level.

APPENDIX D

TASK-CENTERED RATING RUBRIC

Task #1. Colloquial Communication (L11)

L11. Listen to American pilots or air traffic controllers' colloquial (plain) English

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

Task #2. Changed Flight Plan (L5, W4, 5)

L5. Listen to changed flight plan

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

W4. Type the changed flight plan using a computer

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

W5. Take notes of changed flight plan

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

Task #3. Terminal Information (R5, L2, S9, S6)

R5. Read and understand weather information symbols & abbreviation

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

L2. Listen to a pilot' request of weather information

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

S9. Provide weather information to a pilot

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

S6. Provide departure information to a pilot

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

Task #4. Departure Procedure (L4, S6)

L4. Listen to a pilot's departure request

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

S6. Provide departure information to a pilot

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

Task #5. Transition Procedure (L8, S5)

L8. Listen to a pilot's transition request

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

S5. Provide traffic information to transitioning pilots

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

Task #6. Arrival Procedure (L4, S3)

L4. Listen to a pilot's departure request

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

S3. Provide landing instruction to a pilot

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

Task #7. Arrival & Departure Procedure (L1, L4, S3, S4, S5, W1) with NOTAM report (R1, S14, W6)

L4. Listen to a pilot's landing & departure request

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

S3. Provide landing instruction to pilots

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

S4. Provide holding information when traffic is too crowded

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

S5. Provide traffic information to inbound pilots

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

L1. Listen to PIREP

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

W1. Take notes of PIREP

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

R1. Read and understand NOTAM

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

S14. Provide NOTAM to pilots

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

W6. Take notes of NOTAM

Excellent [ ]	Acceptable [ ]	Unacceptable [ ]
Rationale:		

APPENDIX E

TASK PROMPT

Task Situation #1. Colloquial Communication (L11)

Task:
L11. Listen to American pilots or air traffic controllers' colloquial (plain) English

Prompt:
Listen to colleague's two casual greetings – (1) how he or she is doing today; (2) what his or her weekend plan is – and then respond to the two questions.

Scenario:
A colleague avatar next to the test taker (SL user) asks the above two questions. Then, the test taker tries to answer the two questions respectively. If the creation of colleague avatar is more technologically demanding, then another possible situation of receiving an incoming phone call from a colleague and asking the two questions would be also fine.

Functions:
Two pre-recorded audio files which ask the two above questions need to be embedded. Or, the colleague avatar asks the questions. The platform needs to record the test taker's two answers as audio files for rating. Would it possible to integrate a Text-to-speech tool so that I can type in and have the test taker hear the sound during the simulation? This could replace playing pre-recorded audio files.

Sequence:

Q 1. How long have you served as an air traffic controller?

Q 2. How do you like your air traffic control job?

Q3. What is your favorite Korean food? Can you describe it?

Q4. Where can I try the food you like around the base?

Task Situation #2. Changed Flight Plan (L5, W4, 5)

Task:
L5. Listen to changed flight plan; W4. Type the changed flight plan using a computer; W5. Take notes of changed flight plan

Prompt:
Listen to changed flight plan from a grounded pilot in Unicorn 844 (UC 844) and update the changed flight plan in the online flight strip.

Scenario:
A pilot of UC 844, aircraft type CH-47, is reporting their changed flight plan on the ground through radio transmission. The test taker is asked to take note of the flight plan and type the new flight plan in the online flight strip using a keyboard.

Functions:
Prerecorded audio file of native English speaking pilot's radio transmission. The platform needs to record the test taker's utterance as an audio file for rating.

Sequence:

Pilot 1: 'Carol ground', 'UC (Unicorn) 844' on ramp. Request start engine.

(** ATC 1 found UC 844 did not submit a flight plan. Reject their start engine request and ask for a flight plan.)

ATC 1: UC 844 Carol ground, Engine start approved, Altimeter 2995.

_____

Pilot 1: UC 844, would like to amend a flight plan.

ATC 1: UC 844, Go ahead._____

Pilot 1: UC 844, departing R (romeo)-523 at 0900 (nine o'clock), direct to G (golf)-538, 40 (four zero) enroute 01+30 (one hour and thirty minute) ground, and come back here at 1110 (eleven ten) over

ATC 1: UC 844, Roger good copy._____

Task Situation #3. Terminal Information (R5, L2, S9, S6)

Task:
R5. Read and understand weather information symbols & abbreviation; L2. Listen to a pilot' request of weather information; S9. Provide weather information to a pilot; S6. Provide departure information to a pilot

Prompt:
Listen to a pilot's request for terminal information and check the necessary information from the popup window in the screen. Then, provide the information to the pilot.

Scenario:
A pilot of UC 844 requests terminal information for departure. The test taker checks the altimeter, runway, and weather information through a popup window (or a kind of information board) in the screen. After checking, the test taker provides the information to the pilot.

Functions:
Display board shows terminal information as follows: "Daegu weather, estimated ceiling 50 broken visibility 3 mile smoke, runway 02 wind 060 at 15 altimeter 29.95". The platform needs to record the test taker's utterance as an audio file for rating.

* **Current terminal** - Daegu city (R-523): Runway 02, altimeter 29.95

* **Weather**: ceiling 50 broken, visibility 3 mile smoke, wind 060 at 15

Sequence:

Pilot 1: Carol ground, UC (Unicorn) 844, VFR to G (golf)-538, Request Terminal information

ATC 1:

Task Situation #4. Departure Procedure (L4, S6)

Task:
L4. Listen to a pilot's departure request; S6. Provide departure information to a pilot

Prompt:
Listen to a pilot's request for taxi and departure and provide taxi and runway information and departure clearance to the pilot.

Scenario:
A pilot of UC 844 requests terminal information for departure. The test taker checks the altimeter, runway, and weather information through a popup window (or a kind of information board) in the screen. After checking, the test taker provides the information to the pilot.

Functions:
Display board shows terminal information as follows: "Daegu weather, estimated ceiling 50 broken visibility 3 mile smoke, runway 02 wind 060 at 15 altimeter 29.95". The platform needs to record the test taker's utterance as an audio file for rating.

* **Current terminal** - Daegu city (R-523): Runway 02, altimeter 29.95

* **Weather**: ceiling 50 broken, visibility 3 mile smoke, wind 060 at 15

Sequence:
Pilot 1: Carol ground, UC 844 on ramp request taxi instruction for take-off over.
ATC 1: UC 844, Carol ground, Taxi to parallel taxiway via taxiway A and B.
_____
Pilot 1: UC 844 roger. Taxiway A (alpha) and B (bravo).
(UC 844 moves to runway 02 on the screen.)
Pilot 1: Carol ground, UC 844 hover check completed, ready for take-off
ATC 1: UC 844, Hold short of runway 02, Contact tower (121.85)._____
Pilot 1: UC 844 roger. Contact tower
Pilot 1: Carol tower, UC 844 holding short. Ready for take-off.
ATC 1: UC 844, Carol tower, Line up runway 02._____
Pilot 1: Line up runway 02 (zero two) UC 844.
ATC 1: UC 844, Wind 060 at 15, Cleared for take-off. _____
Pilot 1: Cleared for take-off UC 844.
Pilot 1: UC 844, Request upwind departure to East bound.
ATC 1: UC 844 Upwind departure approved, Report check point East.
_____
Pilot 1: UC 844 Roger.
(UC 844 flies to the east bound, the right side and disappears.)
Pilot 1: UC 844 leaving your control zone, request frequency change over.
ATC 1: UC 844, frequency change approved, Have a nice day!!_____
Pilot 1: UC 844 roger

Task Situation #5. Transition Procedure (L8, S5)

Task: L8. Listen to a pilot's transition request; S5. Provide traffic information to transitioning pilots
Prompt:
Listen to a pilot's request for transition and provide traffic information and clearance to the pilot.

Scenario:
A pilot of SP 035 requests traffic information for transition from the east to the west bound. The test taker provides the traffic information to the pilot including the outbound of UC 844.

Functions:
The platform needs to record the test taker's utterance as an audio file for rating.

Sequence:

Pilot 2: Carol tower, SP (Spider) 035 (zero three five) over

ATC 1: SP 035, Carol tower, Go ahead._____

Pilot 2: SP 035, 15 miles East of R (romeo)-523, request transition your control zone east to west at 2000 (two thousand), over.

ATC 1: SP 035 Transition east to west approved, Altimeter 2995. Report leaving my control zone (or Report clear)._____

Pilot 2: SP 035 Roger. Altimeter 29.95 (two niner niner five).

(After SP 035 passing)

(** Warn SP 035 about UC 844 (one Ch-47) which is east bound, 3 miles ahead of them.)

ATC 1: SP 035, Traffic one CH-47 3miles ahead of you._____

Pilot 2: SP 035 Looking for.

Pilot 2: SP 035 traffic insight. Lower altitude than me.

ATC 1: SP 035, Roger maintain visual separation._____

Pilot 2: SP 035 roger. Maintain visual separation.

Task Situation #6. Arrival Procedure (L4, S3)

Task:
L4. Listen to a pilot's departure request; S3. Provide landing instruction to a pilot
Prompt:
Listen to a pilot's request for landing and provide landing instruction to the pilot.
Scenario:
A pilot of SP 313 requests terminal information for landing. The test taker checks the altimeter, runway, and weather information through a popup window (or a kind of information board) in the screen. After checking, the test taker provides the information to the pilot.

Functions:
Display board shows terminal information as follows: "Daegu weather, estimated ceiling 50 broken visibility 3 mile smoke, runway 02 wind 060 at 15 altimeter 29.95". The platform needs to record the test taker's utterance as an audio file for rating.

* **Current terminal** - Daegu city (R-523): Runway 02, altimeter 29.95

* **Weather**: ceiling 50 broken, visibility 3 mile smoke, wind 060 at 15

Sequence:

Pilot 3: Carol tower, SP (Spider) 313 over

ATC 1: SP 313, Carol tower, Go ahead._____

Pilot 3: SP 313, 10 miles east of R (romeo)-523, code 7 on board, request landing over.

ATC 1: SP 313, Runway 02, Altimeter 2995, Report 3miles on final.

_____

Pilot 3: SP 313 Roger, Runway 02, Altimeter 29.95 (two niner niner five)

Pilot 3: SP 313, 3 miles on final.

ATC 1: SP 313, Runway 02, Wind 060 at 15, cleared to land.

_____

(SP 313 is slowly approaching from the east to land on a runway.)

Pilot 3: Cleared to land, Runway 02 (zero two) SP 313.

ATC 1: SP 313, Contact ground (124.0).

Pilot 3: Contact ground 124.0 (one two four point zero), SP 313.

Pilot 3: Carol ground, SP 313 Request taxi to the ramp.

ATC 1: SP 313, Carol ground, Taxi to ramp via taxiway A and B.

_____

Pilot 3: SP 313 taxi to the ramp via taxiway A (alpha) and B (bravo).

Task Situation#7. Arrival & Departure Procedure (L1, L4, S3, S4, S5, W1) with NOTAM report (R1, S14, W6)

Task:
L4. Listen to a pilot's landing & departure request; S3. Provide landing instruction to pilots; S4. Provide holding information when traffic is too heavy; S5. Provide traffic information to inbound pilots; L1. Listen to PIREP; W1. Take notes of PIREP; R1. Read and understand NOTAM; S14. Provide NOTAM to pilots; W6. Take notes of NOTAM

Prompt:
Listen to PIREP and pilot's requests for landing and departure and then provide appropriate responses in timely manner.

Scenario:
Two aircrafts (SP 971, SP 972) are approaching to land providing PIREP (weather report) to ATC. While, UC 823 is going to take off. The test taker is required to respond multiple simultaneous tasks by prioritizing the sequence.

Functions:
Two display boards show (1) terminal information and (2) NOTAM (Notice to Airmen). The platform needs to record the test taker's utterance as an audio file for rating.

* **Current terminal** - Daegu city (R-523): Runway 02, altimeter 29.95

* **Weather**: ceiling 50 broken, visibility 3 mile smoke, wind 060 at 15

* **NOTAM**

Subject: Para Drop

Date: 24 JUL 2014 0500-1200

Sequence:

Pilot 4: Carol ground, UC (Unicorn) 823 on ramp VFR to Chil-gok area, request taxi instruction for take-off, over.
ATC 1: UC 823, Carol tower, Taxi to runway 02 via taxiway A, Hold short of taxiway B.
_____
Pilot 4: UC 823 Roger, runway 02, A (alpha) then hold short of B (bravo).
Pilot 5: Carol tower, SP 971 (niner seven one) flight of two passing Chil-gok area at 23 (twenty three), 2000 (two thousand) encountering turbulence Now 10 miles East of R (romeo)-523, request landing over.
ATC 1: SP 971, Carol tower, Runway 02, Altimeter 2998, Report turning base.
_____
Pilot 5: SP (spider) 971 Roger, altimeter 29.98 (two niner niner eight), Request straight in approach over.
ATC 1: SP 971, Straight in approved, Report 3miles of final.
_____
Pilot 5: SP 971 Roger.
ATC 1: _____
Pilot 5: SP 971, 3 miles on final.
ATC 1: SP 971, Runway 02, wind 060 at 15, cleared to land. _____
Pilot 5: SP 971, roger cleared to land.
Pilot 4: Carol ground, UC 823 Request NOTAM in Chil-gok area.
ATC 1: UC 823, Para drop in progress in Chilgok area at 0500 to 1200.
_____
Pilot 4: UC 823 Roger.
Pilot 5: Carol tower, SP 971 taxi to ramp for full stop.
ATC 1: SP 971, Taxi to ramp approved. _____
Pilot 5: SP 971, roger.
Pilot 4: UC 823, Hover check completed, ready for take-off.
ATC 1: UC 823, Hold position, contact tower. _____
Pilot 4: UC 823, roger contact tower.
Pilot 4: Carol tower, UC 823 Holding short, Ready for take-off.
ATC 1: UC 823 Carol tower, runway 02, wind 060 at 15, Cleared for take-off.
_____
Pilot 4: UC 823, Cleared for take-off.
Pilot 4: Carol tower, UC 823 leaving your control zone, request frequency change over.
ATC 1: UC 823, Frequency change approved. _____
Pilot 4: UC 823 roger.

APPENDIX F

SEMI-STRUCTURED INTERVIEW QUESTIONNAIRE FOR EXPERT AIR TRAFFIC CONTROLLERS

[Domain Description – Research Question 1.1]

1. What specific language skills are needed for successful aviation English communication in the Army Aviation context?

2. What specific language knowledge is needed for successful aviation English communication in the Army Aviation context?

3. What specific language processes are needed for successful aviation English communication in the Army Aviation context?

[Domain Description – Research Question 1.2]

4. What are the possible aviation English assessment tasks that can be representative of the target domain?

5. How would you design aviation English assessment in a virtual environment?

[Domain Description – Research Question 1.3]

6. How effectively does the VITAEA integrate critical aviation English communication skills, knowledge, and processes in the assessment?

7. What additional features needs to be included in the VITAEA for authentic aviation English assessment?

[Evaluation – Research Question 2.1]

8. What criteria or aspect of construct-centered rating rubric needs to be considered?

9. What criteria or aspect of task-centered rating rubric needs to be considered?

10. What is your opinion about the appropriateness of the both construct-centered and task-centered rating rubrics?

[Evaluation – Research Question 2.2]

11. To what extent was task administration condition appropriate for providing evidence of target aviation English ability?

12. What is your opinion about the prompts and task taking process in the VITAEA?

[Evaluation – Research Question 2.3]

13. To what extent was the rating calibration session prior to the actual rating session helpful for your rating practice?

14. Do you have any suggestions for improvement or modification about the rater training?

APPENDIX G

SEMI-STRUCTURED POST-TEST INTERVIEW QUESTIONNAIRE FOR TEST TAKERS

1. How well were you able to be emerged during the tasks?

2. How authentic were the characters and situations of the tasks compared what you learned about ATC?

3. How efficiently were you able to perform ATC in the given tasks?

4. Compared to paper & pencil tests, which one seems to be more reliable and valid test?

5. Overall, to what extent were you satisfied?

6. Any comments on the interface, color, character, menu, mouse movement of the Second Life task environment?

7. To what extent does this SL test seem to be useful in aviation English assessment in the future?

8. Any areas that need to be revised?

9. Any features or contents that need to be added?

APPENDIX H

COMMUNICATION STRATEGY CODING SCHEME

Responding	Message Abandonment [MA]	Leaving a message unfinished because of some language difficulty
	Message Reduction [MR]	Reducing the message by avoiding certain language structures or topics considered problematic language-wise
	Message Replacement [MP]	Substituting the original message with a new one because of not feeling capable of executing it
	Circumlocution (paraphrase) [CL]	Exemplifying or describing the properties of the target object or action (e.g., it became water instead of "melt")
	Restructuring [RS]	Abandoning the execution of a verbal plan because of language difficulties, leaving the utterance unfinished (e.g., On Mickey's face we can see the… so he's he's he's wondering.)
	Literal Translation (transfer) [LT]	Translating literally a lexical item, an idiom, a compound word or structure from L1
	Mumbling [MB]	Muttering inaudibly a word whose correct form the speaker is uncertain about
	Omission [OM]	Leaving a gap when not knowing a word and carrying on as if it had been said
	Retrieval [RT]	In an attempt to retrieve a lexical item saying a series of incomplete or wrong forms before reaching the optimal form. (e.g., It's brake er… it's broken broked broke.)
	Self-Repair [SR]	Making self-initiated corrections in one's own speech
	Use of Filler [UF]	Using gambits to fill pauses, to stall, and to gain time in order to keep the communication channel open
	Self-Repetition [ST]	Repeating a word or a string of words immediately after they were said
	Feigning Understanding [FU]	Making an attempt to carry on the conversation in spite of not understanding something by pretending to understand
	Direct Appeal for Help [DA]	Turning to the examiner for assistance by asking an explicit question concerning a gap in one's L2 knowledge
	Indirect Appeal for Help [IA]	Trying to elicit help from the interlocutor indirectly by expressing lack of a needed L2 item either verbally or nonverbally
	Asking for Repetition [AR]	Requesting repetition when not hearing or understanding something properly (e.g., Pardon? What?)
	Asking for Clarification [AC]	Requesting explanation of an unfamiliar meaning structure
	Expressing Non-understanding [EN]	Expressing that one did not understand something properly either verbally or nonverbally
	Response Repeat [RR]	Repeating the original trigger or the previous phraseology from pilots
	Response Rephrase [RP]	Rephrasing the original trigger or the previous phraseology from pilots